Check for updates

NIST Trustworthy and Responsible AI
NIST AI 100-2e2025

Adversarial Machine Learning

A Taxonomy and Terminology of Attacks and Mitigations

Apostol Vassilev
Alina Oprea
Alie Fordyce
Hyrum Anderson
Xander Davies
Maia Hamin

This publication is available free of charge from:
https://doi.org/10.6028/NIST.AI.100-2e2025

NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

Check for updates

NIST 可信与负责任的人工智能
NIST AI 100-2e2025

对抗性机器学习

攻击与缓解的分类及术语

阿波斯托尔·瓦西列夫
阿丽娜·奥普雷亚 阿里·福代斯
海勒姆·安德森
赞德·戴维斯
玛雅·哈明

本出版物可从以下网址免费获取：
https://doi.org/10.6028/NIST.AI.100-2e2025

NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

本译本由机器翻译生成，供快速理解条文与内部培训使用；
涉及合规判断或法律适用时，请咨询专业顾问，并以官方英文文本为准。
翻译您自己的文件，请选择桐鸾译馆 · https://www.ivyivy.com.cn
扫码关注服务号

NIST Trustworthy and Responsible AI
NIST AI 100-2e2025

Adversarial Machine Learning

A Taxonomy and Terminology of Attacks and Mitigations

Apostol Vassilev
Computer Security Division
Information Technology Laboratory

Alina Oprea
Northeastern University

Maia Hamin
U.S. AI Safety Institute
National Institute of Standards and Technology

Alie Fordyce
Hyrum Anderson
Cisco

Xander Davies
U.K. AI Security Institute

This publication is available free of charge from:
https://doi.org/10.6028/NIST.AI.100-2e2025

March 2025

U.S. Department of Commerce
Howard Lutnick, Secretary

National Institute of Standards and Technology
Craig Burkhardt, Acting Under Secretary of Commerce for Standards and Technology and Acting NIST Director

NIST 可信与负责任的人工智能
NIST AI 100-2e2025

对抗性机器学习

攻击与缓解的分类及术语

阿波斯托尔·瓦西列夫
计算机安全部
信息技术实验室

阿丽娜·奥普雷
东北大学

玛雅·哈明
美国人工智能安全研究所
美国国家标准与技术研究院

阿莉·福代斯
海勒姆·安德森
思科
技术

赞德·戴维斯
英国人工智能安全研究所

本出版物可从以下网址免费获取：
https://doi.org/10.6028/NIST.AI.100-2e2025

2025年3月

U.S 商务部*Howard Lutnick*，部长

美国国家标准与技术研究院
Craig Burkhardt，代理商务部标准与技术副部长兼代理*NIST*主任

Certain commercial equipment, instruments, software, or materials, commercial or non-commercial, are identified in this paper in order to specify the experimental procedure adequately. Such identification does not imply recommendation or endorsement of any product or service by NIST, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

本文中标识了某些商业设备、仪器、软件或材料(商业或非商业)，以便充分说明实验过程。此类标识并不意味着NIST对任何产品或服务的推荐或认可，也不意味着所标识的材料或设备必然是用于该目的的最佳可用选择。

NIST Technical Series Policies
Copyright, Use, and Licensing Statements
NIST Technical Series Publication Identifier Syntax

NIST 技术系列政策
版权、使用和许可声明
NIST 技术系列出版物标识符语法

Publication History
Approved by the NIST Editorial Review Board on 2025-03-20

出版历史
2025-03-20 经 NIST 编辑审查委员会批准

How to Cite this NIST Technical Series Publication:
Vassilev A, Oprea A, Fordyce A, Anderson H, Davies X, Hamin M (2025) Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. (National Institute of Standards and Technology, Gaithersburg, MD) NIST Trustworthy and Responsible AI, NIST AI 100-2e2025.
https://doi.org/10.6028/NIST.AI.100-2e2025

如何引用本 NIST 技术系列出版物：
Vassilev A, Oprea A, Fordyce A, Anderson H, Davies X, Hamin M(2025) 对抗性机器学习：攻击与缓解分类及术语。(美国国家标准与技术研究院，马里兰州盖瑟斯堡)NIST 可信与负责任 AI,NIST AI 100-2e2025。
https://doi.org/10.6028/NIST.AI.100-2e2025

Author ORCID iDs
Apostol Vassilev: 0000-0002-9081-3042
Alina Oprea: 0000-0002-4979-5292
Maia Hamin: 0009-0009-3834-6553

作者 ORCID iD
阿波斯托尔·瓦西列夫：0000-0002-9081-3042
Alina Oprea：0000-0002-4979-5292
Maia Hamin：0009-0009-3834-6553

Contact Information
ai-100-2@nist.gov

联系信息
ai-100-2@nist.gov

Additional Information
Additional information about this publication is available at https://csrc.nist.gov/pubs/ai/100/2/e2025/final, including related content, potential updates, and document history.

附加信息
有关本出版物的更多信息，请访问https://csrc.nist.gov/pubs/ai/100/2/e2025/final,，包括相关内容、可能的更新和文档历史。

All comments are subject to release under the Freedom of Information Act (FOIA).

所有评论均可能根据《信息自由法》(FOIA)予以公开。

Abstract

This NIST Trustworthy and Responsible AI report provides a taxonomy of concepts and defines terminology in the field of adversarial machine learning (AML). The taxonomy is arranged in a conceptual hierarchy that includes key types of ML methods, life cycle stages of attack, and attacker goals, objectives, capabilities, and knowledge. This report also identifies current challenges in the life cycle of AI systems and describes corresponding methods for mitigating and managing the consequences of those attacks. The terminology used in this report is consistent with the literature on AML and is complemented by a glossary of key terms associated with the security of AI systems. Taken together, the taxonomy and terminology are meant to inform other standards and future practice guides for assessing and managing the security of AI systems by establishing a common language for the rapidly developing AML landscape.

Keywords

artificial intelligence; machine learning; attack taxonomy; abuse; data poisoning; evasion; privacy breach; attack mitigation; large language model; chatbot.

NIST Trustworthy and Responsible AI

The National Institute of Standards and Technology (NIST) promotes U.S. innovation and industrial competitiveness by advancing measurement science, standards, and technology in ways that enhance economic security and improve our quality of life. Among its broad range of activities, NIST contributes to the research, standards, evaluations, and data required to advance the development, use, and assurance of trustworthy artificial intelligence (AI).

摘要

本NIST可信与负责任人工智能报告提供了对抗性机器学习(AML)领域的概念分类法并定义了术语。该分类法按概念层级编排，涵盖机器学习方法的关键类型、攻击的生命周期阶段以及攻击者的目标、目的、能力和知识。本报告还识别了人工智能系统生命周期中的当前挑战，并描述了缓解和管理这些攻击后果的相应方法。本报告使用的术语与AML文献一致，并辅以与人工智能系统安全相关的关键术语表。总体而言，该分类法和术语旨在通过为快速发展的AML格局建立共同语言，为其他标准和未来用于评估和管理人工智能系统安全的实践指南提供参考。

关键词

人工智能；机器学习；攻击分类法；滥用；数据投毒；规避；隐私泄露；攻击缓解；大语言模型；聊天机器人。

NIST 可信与负责任的人工智能

美国国家标准与技术研究院(NIST)通过推进测量科学、标准和技术的进步来促进美国创新和产业竞争力，从而增强经济安全并改善我们的生活质量。在其广泛的活动范围内，NIST为推进可信人工智能(AI)的开发、使用和保障所需的研究、标准、评估和数据作出贡献。

Table of Contents

Audience . . . viii
Background . . . viii
Trademark Information . . . viii
How to Read This Document . . . ix
Acknowledgments . . . ix
Author Contributions . . . ix
Predictive AI and Generative AI Taxonomy Index . . . x
Executive Summary . . . xii
1. Introduction . . . 1
2. Predictive AI Taxonomy . . . 4
2.1. Attack Classification . . . 4
2.1.1. Stages of Learning . . . 5
2.1.2. Attacker Goals and Objectives . . . 6
2.1.3. Attacker Capabilities . . . 7
2.1.4. Attacker Knowledge . . . 8
2.1.5. Data Modality . . . 9
2.2. Evasion Attacks and Mitigations . . . 11
2.2.1. White-Box Evasion Attacks . . . 12
2.2.2. Black-Box Evasion Attacks . . . 15
2.2.3. Transferability of Attacks . . . 15
2.2.4. Evasion attacks in the real world . . . 16
2.2.5. Mitigations . . . 17
2.3. Poisoning Attacks and Mitigations . . . 19
2.3.1. Availability Poisoning . . . 19
2.3.2. Targeted Poisoning . . . 21
2.3.3. Backdoor Poisoning . . . 22
2.3.4. Model Poisoning . . . 26
2.3.5. Poisoning Attacks in the Real World . . . 27
2.4. Privacy Attacks and Mitigations . . . 28
2.4.1. Data Reconstruction . . . 28

目录

受众 . . . viii
背景 . . . viii
商标信息 . . . viii
如何阅读本文档 . . . ix 致谢 . . . ix 作者贡献 . . . ix

预测式AI与生成式AI分类法索引 . . . x

执行摘要 . . . xii
1. 引言 . . . 1
2. 预测式人工智能分类法 . . . 4
2.1. 攻击分类 . . . 4
2.1.1. 学习阶段 . . . 5
2.1.2. 攻击者的目标与目的 . . . 6
2.1.3. 攻击者能力 . . . 7
2.1.4. 攻击者知识 . . . 8
2.1.5. 数据模态 . . . 9
2.2. 规避攻击与缓解 . . . 11
2.2.1. 白盒规避攻击 . . . 12
2.2.2. 黑盒规避攻击 . . . 15
2.2.3. 攻击的可迁移性 . . . 15
2.2.4. 现实世界中的规避攻击 . . . 16
2.2.5. 缓解措施 . . . 17
2.3. 投毒攻击与缓解 . . . 19
2.3.1. 可用性投毒 . . . 19
2.3.2. 定向投毒 . . . 21
2.3.3. 后门投毒 . . . 22
2.3.4. 模型投毒 . . . 26
2.3.5. 现实世界中的投毒攻击 . . . 27
2.4. 隐私攻击与缓解 . . . 28
2.4.1. 数据重建 . . . 28

2.4.2. Membership Inference 29
2.4.3. Property Inference 30
2.4.4. Model Extraction 31
2.4.5. Mitigations 32
3. Generative AI Taxonomy 34
3.1. Attack Classification 34
3.1.1. GenAI Stages of Learning 36
3.1.2. Attacker Goals and Objectives 39
3.1.3. Attacker Capabilities 40
3.2. Supply Chain Attacks and Mitigations 41
3.2.1. Data Poisoning Attacks 42
3.2.2. Model Poisoning Attacks 42
3.2.3. Mitigations 42
3.3. Direct Prompting Attacks and Mitigations 43
3.3.1. Attack Techniques 44
3.3.2. Information Extraction 46
3.3.3. Mitigations 48
3.4. Indirect Prompt Injection Attacks and Mitigations 50
3.4.1. Availability Attacks 51
3.4.2. Integrity Attacks 51
3.4.3. Privacy Compromise 52
3.4.4. Mitigations 53
3.5. Security of Agents 54
3.6. Benchmarks for AML Vulnerabilities 54
4. Key Challenges and Discussion 55
4.1. Key Challenges in AML 55
4.1.1. Trade-Offs Between the Attributes of Trustworthy AI 55
4.1.2. Theoretical Limitations on Adversarial Robustness 56
4.1.3. Evaluation 57
4.2. Discussion 57
4.2.1. The Scale Challenge 57

2.4.2. 成员推断 29
2.4.3. 属性推断 30
2.4.4. 模型萃取 31
2.4.5. 缓解措施 32
3. 生成式人工智能分类法 34
3.1. 攻击分类 34
3.1.1. GenAI 学习阶段 36
3.1.2. 攻击者的目标与目的 39
3.1.3. 攻击者能力 40
3.2. 供应链攻击与缓解措施 41
3.2.1. 数据投毒攻击 42
3.2.2. 模型投毒(poisoning)攻击 42
3.2.3. 缓解措施 42
3.3. 直接提示攻击与缓解 43
3.3.1. 攻击技术 44
3.3.2. 信息提取 46
3.3.3. 缓解措施 48
3.4. 间接提示注入攻击与缓解 50
3.4.1. 可用性攻击 51
3.4.2. 完整性攻击 51
3.4.3. 隐私泄露 52
3.4.4. 缓解措施 53
3.5. 智能体安全 54
3.6. AML 漏洞的基准测试 54
4. 关键挑战与讨论 55
4.1. AML 中的关键挑战 55
4.1.1. 可信AI各属性之间的权衡 55
4.1.2. 对抗鲁棒性的理论局限性 56
4.1.3. 评估 57
4.2. 讨论 57
4.2.1. 规模挑战 57

4.2.2. Supply Chain Challenges 58
4.2.3. Multimodal Models 58
4.2.4. Quantized Models 59
4.2.5. Risk Management in Light of AML 59
4.2.6. AML and Other AI System Characteristics 60
Appendix: Glossary 107

List of Figures

Figure 1. **Taxonomy of attacks on PredAI systems** 4
Figure 2. **Taxonomy of attacks on GenAI systems** 35
Figure 3. **Example LLM Training Pipeline used for InstructGPT [281]** 36
Figure 4. **LLM enterprise adoption pipeline** 37
Figure 5. **LLM enterprise adoption reference architecture** 38
Figure 6. **Retrieval-augmented generation** 39
Figure 7. **Map of the development and deployment life cycle of an AI model for broad-scale query access** 47
Figure 8. Pareto optimality 55

4.2.2. 供应链挑战 58
4.2.3. 多模态模型 58
4.2.4. 量化模型 59
4.2.5. 基于AML的风险管理 59
4.2.6. AML 及其他 AI 系统特征.....60
附录：术语表 107

图目录

图1. **针对PredAI系统的攻击分类** 4
图2. **针对GenAI系统的攻击分类** 35
图3. **用于InstructGPT的示例LLM训练流程[281]** 36
图 4. **LLM 企业采用流水线** 37
图 5. LLM 企业采用参考架构 38
图 6. **检索增强生成**..... 39
图7. **AI模型开发与部署生命周期映射用于大规模查询访问**..... 47
图8. 帕累托最优性 55

Audience

The intended primary audience for this document includes individuals and groups who are responsible for designing, developing, deploying, evaluating, and governing AI systems.

Background

This document is the result of an extensive literature review, conversations with experts in adversarial machine learning, and research performed by the authors in adversarial machine learning.

Trademark Information

All trademarks and registered trademarks belong to their respective organizations.

The Information Technology Laboratory (ITL) at NIST develops tests, test methods, reference data, proof of concept implementations, and technical analyses to advance the development and productive use of information technology. ITL's responsibilities include the development of management, administrative, technical, and physical standards and guidelines.

This NIST Trustworthy and Responsible AI report focuses on identifying, addressing, and managing risks associated with adversarial machine learning. While practical guidance[1] published by NIST may serve as an informative reference, this guidance remains voluntary.

The content of this document reflects recommended practices. This document is not intended to serve as or supersede existing regulations, laws, or other mandatory guidance.

[1] In the context of this paper, the terms "practice guide," "guide," "guidance," and the like are consensus-created informative references that are intended for voluntary use. They should not be interpreted as equal to the use of the term "guidance" in a legal or regulatory context. This document does not establish any legal standard or any other legal requirement or defense under any law, nor does it have the force or effect of law.

受众

本文档的预期主要受众包括负责设计、开发、部署、评估和治理人工智能系统的个人和团体。

背景

本文档是广泛文献综述、与对抗性机器学习专家的交流以及作者在对抗性机器学习领域所开展研究的结果。

商标信息

所有商标和注册商标均归其各自组织所有。

NIST 的信息技术实验室(ITL)开发测试、测试方法、参考数据、概念验证实现和技术分析，以促进信息技术的发展和高效利用。ITL 的职责包括制定管理、行政、技术和物理标准及指南。

本 NIST 可信与负责任人工智能报告侧重于识别、应对和管理与对抗性机器学习相关的风险。虽然 NIST 发布的实用指南[1]可作为参考信息，但该指南仍属自愿性质。

本文件内容反映推荐做法。本文件无意作为或取代现有法规、法律或其他强制性指南。

1 在本文语境中，“实践指南”“指南”“指导”等术语是经共识形成的信息性参考，旨在供自愿使用。它们不应被解释为等同于法律或监管语境中“指导”一词的使用。本文件不确立任何法律标准或任何其他法律要求或抗辩，也不具有法律效力。

How to Read This Document

This document uses the terms "AI technology," "AI system," and "AI applications" interchangeably. Terms related to the machine learning pipeline, such as "ML model" or "algorithm," are also used interchangeably in this document. Depending on context, the term "system" may refer to the broader organizational and/or social ecosystem within which the technology was designed, developed, deployed, and used instead of the more traditional use related to computational hardware or software.

Important reading notes:

- This document includes a series of blue callout boxes that highlight nuances and important takeaways.
- This document contains links shown in blue. Clicking on them will bring the reader to the relevant resource. Links links in the References point to external sources.
- Terms that are used but not defined or explained in the text are listed and defined in the Glossary. They are displayed in small caps in the text. Clicking on a word shown in SMALL CAPS (e.g., ADVERSARIAL EXAMPLE) takes the reader directly to the definition of that term in the Glossary. From there, one may click on the page number shown at the end of the definition to return.
- This document provides an Index of attack types to easily navigate and reference attacks and corresponding mitigations.

Acknowledgments

The authors wish to thank all of the people and organizations who submitted comments on the draft version of this paper. The received comments and suggested references were essential to improving the document and the future direction of this work. The authors also want to thank the many NIST, U.S. AI Safety Institute, and U.K. AI Security Institute colleagues who assisted in updating this document.

Author Contributions

The authors contributed equally to this work.

如何阅读本文档

本文件交替使用"人工智能技术""人工智能系统"和"人工智能应用"等术语。与机器学习流水线相关的术语，如"ML 模型"或"算法"，在本文件中亦交替使用。视语境而定，"系统"一词可能指技术被设计、开发、部署和使用所处的更广泛的组织和/或社会生态系统，而非与计算硬件或软件相关的更传统用法。

重要阅读说明：

- 本文件包含一系列蓝色标注框，用于突出细微差别和重要要点。
- 本文件包含以蓝色显示的链接。点击链接将把读者带到相关资源。参考文献中的链接指向外部来源。
- 文中使用但未在正文中定义或解释的术语列于术语表中并予以定义。它们在文中以小型大写字母显示。点击以小型大写字母显示的词语(例如 ADVERSARIAL EXAMPLE)会直接将读者带到术语表中该术语的定义处。从那里，可以点击定义末尾显示的页码返回。
- 本文件提供攻击类型索引，便于浏览和查阅攻击及相应缓解措施。

致谢

作者感谢所有对本文件草案版本提交评论的个人和组织。收到的评论和建议的参考文献对于改进本文件及本工作的未来方向至关重要。作者还感谢许多 NIST、美国人工智能安全研究所和英国人工智能安全研究所的同事协助更新本文件。

作者贡献

作者对本工作贡献相同。

Predictive AI and Generative AI Taxonomy Index

- **Predictive AI Attacks Taxonomy**
  - Availability Violations (ID: NISTAML.01)
    - Model Poisoning (ID: NISTAML.011)
    - Clean-label Poisoning (ID: NISTAML.012)
    - Data Poisoning (ID: NISTAML.013)
    - Energy-latency (ID: NISTAML.014)
  - Integrity Violations (ID: NISTAML.02)
    - Clean-label Poisoning (ID: NISTAML.012)
    - Clean-label Backdoor (ID: NISTAML.021)
    - Evasion (ID: NISTAML.022)
    - Backdoor Poisoning (ID: NISTAML.023)
    - Targeted Poisoning (ID: NISTAML.024)
    - Black-box Evasion (ID: NISTAML.025)
    - Model Poisoning (ID: NISTAML.026)
  - Privacy Compromises (ID: NISTAML.03)
    - Model Extraction (ID: NISTAML.031)
    - Reconstruction (ID: NISTAML.032)
    - Membership Inference (ID: NISTAML.033)
    - Property Inference (ID: NISTAML.034)
  - Supply Chain Attacks (ID: NISTAML.05)
    - Model Poisoning (ID: NISTAML.051)
- **Generative AI Attacks Taxonomy**
  - Availability Violations (ID: NISTAML.01)
    - Data Poisoning (ID: NISTAML.013)
    - Indirect Prompt Injection (ID: NISTAML.015)
    - Prompt Injection (ID: NISTAML.018)
  - Integrity Violations (ID: NISTAML.02)

预测性 AI 与生成式 AI 分类索引

• 预测性 AI 攻击分类

-可用性违规(ID: NISTAML.01)

⋆ 模型投毒(ID:NISTAML.011) ⋆ 干净标签投毒(ID:NISTAML.012) ⋆ 数据投毒(ID:NISTAML.013) ⋆ 能量-延迟(ID:NISTAML.014)

-完整性违规(ID: NISTAML.02)

⋆ 干净标签投毒(ID:NISTAML.012) ⋆ 干净标签后门(ID:NISTAML.021) ⋆ 规避(ID:NISTAML.022)

⋆ 后门投毒(ID:NISTAML.023) ⋆ 定向投毒(ID:NISTAML.024) ⋆ 黑盒规避(ID:NISTAML.025) ⋆ 模型投毒(ID:NISTAML.026)-隐私泄露(ID: NISTAML.03)

⋆ 模型提取(ID:NISTAML.031) ⋆ 重建(ID:NISTAML.032)

⋆ 成员推断(ID:NISTAML.033) ⋆ 属性推断(ID:NISTAML.034)

-供应链攻击(ID: NISTAML.05)

⋆ 模型投毒(ID:NISTAML.051)

• **生成式 AI 攻击分类法**

-可用性违规(ID: NISTAML.01)

⋆ 数据投毒(ID:NISTAML.013)

⋆ 间接提示注入(ID:NISTAML.015) ⋆ 提示注入(ID:NISTAML.018)

-完整性违规(ID: NISTAML.02)

- Data Poisoning (ID: NISTAML.013)
- Indirect Prompt Injection (ID: NISTAML.015)
- Prompt Injection (ID: NISTAML.018)
- Backdoor Poisoning (ID: NISTAML.023)
- Targeted Poisoning (ID: NISTAML.024)
- Misaligned Outputs (ID: NISTAML.027)

- Privacy Compromises (ID: NISTAML.03)
  - Indirect Prompt Injection (ID: NISTAML.015)
  - Prompt Injection (ID: NISTAML.018)
  - Backdoor Poisoning (ID: NISTAML.023)
  - Membership Inference (ID: NISTAML.033)
  - Prompt Extraction (ID: NISTAML.035)
  - Leaking information from user interactions (ID: NISTAML.036)
  - Training Data Attacks (ID: NISTAML.037)
  - Data Extraction (ID: NISTAML.038)
  - Compromising connected resources (ID: NISTAML.039)
- Misuse Violations (ID: NISTAML.04)
  - Prompt Injection (ID: NISTAML.018)
- Supply Chain Attacks (ID: NISTAML.05)
  - Model Poisoning (ID: NISTAML.051)

- 数据投毒(ID:NISTAML.013)

- 间接提示注入(ID:NISTAML.015) - 提示注入(ID:NISTAML.018) - 后门投毒(ID:NISTAML.023) - 定向投毒(ID:NISTAML.024) - 失准输出(ID:NISTAML.027)-隐私泄露(ID: NISTAML.03)
  - 间接提示注入(ID:NISTAML.015) - 提示注入(ID:NISTAML.018) - 后门投毒(ID:NISTAML.023) - 成员推断(ID:NISTAML.033) - 提示提取(ID:NISTAML.035)
  - 从用户交互中泄露信息信息来自用户交互(ID:NISTAML.036) - 训练数据攻击(ID:NISTAML.037) - 数据提取(ID:NISTAML.038)
  - 破坏连接的资源(ID:NISTAML.039)-滥用违规(ID:NISTAML.04)
    - 提示注入(ID:NISTAML.018)
-供应链攻击(ID: NISTAML.05)
  - 模型投毒(ID:NISTAML.051)

Executive Summary

This NIST Trustworthy and Responsible AI report describes a taxonomy and terminology for ADVERSARIAL MACHINE LEARNING (AML) that may aid in securing applications of artificial intelligence (AI) against adversarial manipulations and attacks.

The statistical, data-based nature of ML systems opens up new potential vectors for attacks against these systems' security, privacy, and safety, beyond the threats faced by traditional software systems. These challenges span different phases of ML operations such as the potential for adversarial manipulation of training data; the provision of adversarial inputs to adversely affect the performance of the AI system; and even malicious manipulations, modifications, or interactions with models to exfiltrate sensitive information from the model's training data or to which the model has access. Such attacks have been demonstrated under real-world conditions, and their sophistication and impacts have been increasing steadily.

The field of AML is concerned with studying these attacks. It must consider the capabilities of attackers, the model or system properties that attackers might seek to violate in pursuit of their objectives, and the design of attack methods that exploit vulnerabilities during the development, training, and deployment phases of the ML life cycle. It is also concerned with the design of ML algorithms and systems that can withstand these security and privacy challenges, a property often known as robustness [274].

To taxonimize these attacks, this report differentiates between predictive and generative AI systems and the attacks relevant to each. It considers the components of an AI system including the data; the model itself; the processes for training, testing, and deploying the model; and the broader software and system contexts into which models may be embedded, such as cases where Generative Artificial Intelligence (GenAI) models are deployed with access to private data or equipped with tools to take actions with real-world consequences.

Thus, the attacks within this taxonomy are classified relative to: (i) the AI system type, (ii) the stage of the ML life cycle process in which the attack is mounted, (iii) the attacker's goals and objectives in terms of the system properties they seek to violate, (iv) the attacker's capabilities and access, and (v) the attacker's knowledge of the learning process and beyond.

This report adopts the concepts of security, resilience, and robustness of ML systems from the NIST AI Risk Management Framework. Security, resilience, and robustness are gauged by risk, which is a measure of the extent to which an entity (e.g., a system) is threatened by a potential circumstance or event (e.g., an attack) and the severity of the outcome should such an event occur. However, this report does not make recommendations on risk tolerance (i.e., the level of risk that is acceptable to organizations or society) because it is highly contextual and specific to applications and use cases.

执行摘要

本 NIST 可信与负责任人工智能报告描述了对抗性机器学习(AML)的分类法和术语，可有助于保护人工智能(AI)应用免受对抗性操纵和攻击。

机器学习系统基于统计和数据的特性，为针对这些系统安全性、隐私性和安全性的攻击开辟了新的潜在途径，超出了传统软件系统所面临的威胁。这些挑战贯穿机器学习运行的不同阶段，例如对训练数据进行对抗性操纵的可能性；提供对抗性输入以不利影响人工智能系统性能；甚至对模型进行恶意操纵、修改或交互，以从模型的训练数据或模型可访问的数据中窃取敏感信息。此类攻击已在真实世界条件下得到验证，其复杂性和影响一直在稳步增加。

AML领域关注研究这些攻击。它必须考虑攻击者的能力、攻击者在追求其目标时可能试图破坏的模型或系统属性，以及利用机器学习生命周期开发、训练和部署阶段漏洞的攻击方法设计。它还关注能够抵御这些安全与隐私挑战的机器学习算法和系统的设计，这种属性通常被称为鲁棒性[274]。

为了对这些攻击进行分类，本报告区分了预测式人工智能系统与生成式人工智能系统以及各自相关的攻击。报告考察了人工智能系统的各个组成部分，包括数据；模型本身；训练、测试和部署模型的过程；以及模型可能嵌入其中的更广泛的软件和系统环境，例如生成式人工智能(GenAI)模型在可访问私有数据或配备可采取具有现实后果行动的工具的情况下部署的情形。

因此，本分类法中的攻击是相对于以下方面进行分类的：(i)人工智能系统类型，(ii)攻击所发起的机器学习生命周期过程阶段，(iii)攻击者就其试图破坏的系统属性而言的目标和目的，(iv)攻击者的能力和访问权限，以及(v)攻击者对学习过程及之外的了解。

本报告采用 NIST 人工智能风险管理框架中关于机器学习系统安全性、韧性和鲁棒性的概念。安全性、韧性和鲁棒性通过风险来衡量，风险是对某一实体(例如系统)受到潜在情况或事件(例如攻击)威胁的程度以及此类事件发生时后果严重性的度量。然而，本报告不就风险容忍度(即组织或社会可接受的风险水平)提出建议，因为这高度依赖于具体情境且特定于应用和用例。

The spectrum of effective attacks against ML is wide, rapidly evolving, and covers all phases of the ML lifecycle — from design and implementation to training, testing, and deployment in the real world. The nature and power of these attacks are different and their impacts may depend not only on the vulnerabilities of the ML models but also the weaknesses of the infrastructure in which the AI systems are deployed. AI system components may also be adversely affected by design and implementation flaws that cause failures outside the context of adversarial use, such as inaccuracy. However, these kinds of flaws are not within the scope of the literature on AML or the attacks in this report.

In addition to defining a taxonomy of attacks, this report provides corresponding methods for mitigating and managing the consequences of those attacks in the life cycle of AI systems, and outlines the limitations of widely used mitigation techniques to raise awareness and help organizations increase the efficacy of their AI risk-mitigation efforts. The terminology used in this report is consistent with the literature on AML and is complemented by a glossary that defines key terms associated with the field of AML in order to assist non-expert readers. Taken together, the taxonomy and terminology are meant to inform other standards and future practice guides for assessing and managing the security of AI systems by establishing a common language for the rapidly developing AML landscape. Like the taxonomy, the terminology and definitions are not intended to be exhaustive but rather to serve as a starting point for understanding and aligning on key concepts that have emerged in the AML literature.

针对机器学习的有效攻击范围广泛、快速演变，并涵盖机器学习生命周期的所有阶段--从设计与实现到训练、测试和现实世界中的部署。这些攻击的性质和威力各不相同，其影响可能不仅取决于机器学习模型的脆弱性，还取决于部署人工智能系统所依赖的基础设施的弱点。人工智能系统组件还可能受到设计和实现缺陷的不利影响，这些缺陷会导致在非对抗性使用情境下发生故障，例如不准确性。然而，这类缺陷不属于对抗性机器学习文献或本报告所述攻击的范畴。

除定义攻击分类法外，本报告还提供了在人工智能系统生命周期中缓解和管理这些攻击后果的相应方法，并概述了广泛使用的缓解技术的局限性，以提高认识并帮助组织提升其人工智能风险缓解工作的效力。本报告使用的术语与对抗性机器学习文献保持一致，并辅以术语表，定义对抗性机器学习领域的关键术语，以帮助非专业读者。总体而言，分类法和术语旨在通过为快速发展的对抗性机器学习领域建立共同语言，为其他标准和未来用于评估和管理人工智能系统安全的实践指南提供参考。与分类法一样，术语和定义并非意在穷尽，而是作为理解和就对抗性机器学习文献中涌现的关键概念达成一致的起点。

1. Introduction

Artificial intelligence (AI) systems have been on a global expansion trajectory for several years [267]. These systems are being developed by and widely deployed into the economies of numerous countries, with increasing opportunities for people to use AI systems in many spheres of their lives [92]. This report distinguishes between two broad classes of AI systems: predictive AI (PredAI) and generative AI (GenAI). Although the majority of industrial applications of AI systems are still dominated by PredAI systems, there has been a recent increase in the adoption of GenAI systems in business and consumer contexts. As these systems permeate the digital economy and become essential parts of daily life, the need for their secure, robust, and resilient operation grows. These operational attributes are critical elements of trustworthy AI in the NIST AI Risk Management Framework [274] and the NCSC Machine Learning Principles [266].

The field of ADVERSARIAL MACHINE LEARNING (AML) studies attacks against ML systems that exploit the statistical, data-based nature of ML systems. Despite the significant progress of AI and machine learning (ML) in different application domains, these technologies remain vulnerable to attacks that can cause spectacular failures. The chances of these kinds of failure increase as ML systems are used in contexts where they may be subject to novel or adversarial interactions, and the consequences grow more dire as these systems are used in increasingly high-stakes domains. For example, in PredAI computer vision applications for object detection and classification, well-known cases of adversarial perturbations of input images have caused autonomous vehicles to swerve into lanes going in the opposite direction, stop signs to be misclassified as speed limit signs, and even people wearing glasses to be misidentified in high-security settings [121, 187, 332, 349]. Similarly, the potential for adversarial input to trick ML models into revealing hidden information has become more urgent as more ML models are being deployed in fields like medicine, where medical record leaks can expose sensitive personal information [25, 171].

In GenAI, large language models (LLMs) [13, 15, 49, 85, 102, 236, 247, 277, 279, 348, 365, 371, 372, 436] are increasingly becoming an integral part of software applications and internet infrastructure. LLMs are being used to create more powerful online search tools, help software developers write code, and power chatbots that are used by millions of people every day [255]. LLMs are also being augmented to create more useful AI systems, including through interactions with corporate databases and documents to enable powerful RETRIEVAL-AUGMENTED GENERATION (RAG) (RAG) [210] and through training- or inference-time techniques to enable LLMs to take real-world actions, such as browsing the web or using a bash terminal as an LLM-based AGENT [167, 261, 278, 419]. Thus, vulnerabilities in GenAI systems may expose a broad attack surface for threats to the privacy of sensitive user data or proprietary information about models' architecture or training data, and create risks to the integrity and availability of widely used systems.

As GenAI adoption has grown, the increasing capability of these systems has created another challenge for model developers: how to manage the risks created by unwanted or

1. 引言

人工智能(AI)系统多年来一直处于全球扩张轨道[267]。这些系统由众多国家开发并广泛部署到其经济中，人们在生活的许多领域使用人工智能系统的机会日益增多[92]。本报告区分了两大类人工智能系统：预测式人工智能(PredAI)和生成式人工智能(GenAI)。尽管人工智能系统的多数工业应用仍以 PredAI 系统为主，但近期 GenAI 系统在商业和消费场景中的采用有所增加。随着这些系统渗透数字经济并成为日常生活的重要组成部分，对其安全、鲁棒和韧性运行的需求日益增长。这些运行属性是 NIST 人工智能风险管理框架[274]和 NCSC 机器学习原则[266]中可信人工智能的关键要素。

对抗性机器学习(AML)领域研究针对机器学习系统的攻击，这些攻击利用机器学习系统基于数据的统计特性。尽管人工智能和机器学习(ML)在不同应用领域取得了显著进展，但这些技术仍然容易受到可能导致严重故障的攻击。随着机器学习系统被用于可能面临新颖或对抗性交互的场景，此类故障发生的概率随之增加；而这些系统被应用于日益高风险的领域时，后果也愈发严重。例如，在PredAI计算机视觉的物体检测与分类应用中，众所周知的输入图像对抗性扰动案例已导致自动驾驶车辆转向驶入对向车道、停止标志被误分类为限速标志，甚至在高安全环境中佩戴眼镜的人员被错误识别[121,187,332,349]。同样，随着越来越多的机器学习模型被部署到医疗等领域，对抗性输入诱骗机器学习模型泄露隐藏信息的潜在风险变得更加紧迫，因为医疗记录泄露可能暴露敏感的个人信息[25,171]。

在GenAI中，大语言模型(LLM)[13,15,49,85,102,236,247,277,279,348,365,371,372,436]正日益成为软件应用和互联网基础设施不可或缺的组成部分。LLM被用于创建更强大的在线搜索工具、帮助软件开发者编写代码，以及驱动每天被数百万人使用的聊天机器人[255]。LLM还通过多种方式被增强以构建更有用的人工智能系统，包括与企业数据库和文档交互以实现强大的检索增强生成(RAG)[210]，以及通过训练时或推理时技术使LLM能够采取现实世界行动，例如作为基于LLM的智能体浏览网页或使用bash终端[167,261,278,419]。因此，GenAI系统中的漏洞可能为威胁敏感用户数据隐私或模型架构、训练数据等专有信息暴露广泛的攻击面，并对广泛使用系统的完整性和可用性构成风险。

随着GenAI应用的普及，这些系统能力的不断增强给模型开发者带来了另一个挑战：如何管理由不期望的或

harmful uses of these systems' capabilities.[275] As model developers have increasingly sought to apply technical interventions to reduce models' potential for misuse, another surface for high-stakes AML attacks has emerged in attacks that attempt to circumvent or disrupt these protections.

Fundamentally, many AI systems are susceptible both to AML attacks and to attacks that more closely resemble traditional cybersecurity attacks, including attacks against the platforms on which they are deployed. This report focuses on the former and considers the latter to be within the scope of traditional cybersecurity taxonomies.

Both PredAI and GenAI systems are vulnerable to attacks enabled by a range of attacker capabilities throughout the development and deployment life cycle. Attackers can manipulate training data [327], including the Internet data used in large-scale model training [57], or can modify test-time inference data and resources by adding adversarial perturbations or suffixes. Attackers can also attack the components used to make AI systems by inserting TROJAN functionality. As organizations increasingly rely on pre-trained models that could be used directly or fine-tuned with new datasets to enable different tasks, their vulnerability to these attacks increases.

Modern cryptography often relies on algorithms that are secure in an information-theoretic sense, that is, those that can be formally proven to ensure security under certain conditions. However, there are no information-theoretic security proofs for the widely used ML algorithms in modern AI systems. Moreover, information-theoretic *impossibility* results that set limits on the effectiveness of widely used mitigation techniques have begun to appear in the literature [124, 140, 432]. As a result, many of the advances in developing mitigations against different classes of AML attacks tend to be empirical and limited in nature, adopted because they appear to work in practice rather than because they provide information-theoretic security guarantees. Thus, many of these mitigations may themselves be vulnerable to new discoveries and evolutions in attacker techniques.

This report offers guidance for the development of:

- Standardized terminology for AML terms that can be used across relevant ML and cybersecurity communities. There are notable differences in terminology in different stakeholder communities and it is important to work towards bridging the differences as AI is increasingly adopted throughout enterprise and consumer contexts.
- A taxonomy of the most widely studied and currently effective attacks in AML, including:
  - Evasion, poisoning, and privacy attacks for PredAI systems
  - Poisoning, direct prompting, and indirect prompt injection attacks for GenAI systems
- A discussion of potential mitigations for these attacks and the limitations of existing mitigation techniques

有害使用所创造的风险。[275] 随着模型开发者日益寻求采用技术干预措施以降低模型被滥用的可能性，另一个高风险 AML 攻击面已经出现，即试图规避或破坏这些防护措施的攻击。

从根本上说，许多人工智能系统既易受 AML 攻击影响，也易受更接近传统网络安全攻击的攻击影响，包括针对其部署平台的攻击。本报告侧重于前者，并认为后者属于传统网络安全分类法的范围。

PredAI和GenAI系统均易受到在整个开发和部署生命周期中由一系列攻击者能力所支撑的攻击。攻击者可以操纵训练数据[327]，包括大规模模型训练中使用的互联网数据[57]，或者通过添加对抗性扰动或后缀来修改测试时推理数据和资源。攻击者还可以通过插入木马功能来攻击用于构建人工智能系统的组件。随着组织日益依赖可直接使用或利用新数据集微调以实现不同任务的预训练模型，它们对这些攻击的脆弱性也随之增加。

现代密码学通常依赖于在信息论意义上安全的算法，即那些可以被形式化证明在某些条件下确保安全的算法。然而，现代人工智能系统中广泛使用的机器学习算法并不存在信息论安全证明。此外，文献中已开始出现信息论不可能性结果，这些结果对广泛使用的缓解技术的有效性设定了限制[124,140,432]。因此，针对不同类别AML攻击开发缓解措施的许多进展往往是经验性的且具有局限性，采用它们是因为在实践中似乎有效，而非因为它们提供了信息论安全保证。因此，这些缓解措施中的许多本身可能容易受到攻击者技术的新发现和演进的影响。

本报告为以下方面的发展提供指导：

- 用于AML术语的标准化术语，可在相关ML和网络安全社区中通用。
  不同利益相关者社区在术语上存在显著差异，随着AI在企业与消费者场景中的日益普及，弥合这些差异至关重要。
- 对AML中研究最广泛且当前最有效的攻击的分类法，
  包括：
  - PredAI 系统的规避、投毒和隐私攻击
  -针对GenAI系统的投毒、直接提示和间接提示注入攻击系统
- 对这些攻击的潜在缓解措施及现有缓解技术局限性的讨论

NIST intends to update this report as new developments emerge in AML attacks and mitigations.

This report provides a categorization of common classes of attacks and their mitigations for PredAI and GenAI systems. This report is not intended to provide an exhaustive survey of all available literature on Adversarial ML, which includes more than 11,354 references on arXiv.org since 2021 as of July 2024.

This report is organized into three sections.

- Section 2 considers PredAI systems. Section 2.1 introduces the taxonomy of attacks for PredAI systems, which defines the broad categories of attacker objectives and goals, and identifies the capabilities that an adversary must leverage to achieve the corresponding objectives. Specific attack classes are also introduced for each type of capability. Sections 2.2, 2.3, and 2.4 discuss the major classes of attacks: evasion, poisoning, and privacy, respectively. A corresponding set of mitigations for each class of attacks is provided in the attack class sections.
- Section 3 considers GenAI systems. Section 3.1 introduces the taxonomy of attacks for GenAI systems and defines the broad categories of attacker objectives and adversary capabilities relevant to these systems. Specific attack classes are introduced for each type of capability, along with relevant mitigations.
- Section 4 discusses remaining challenges in the field, including limitations to widely used mitigation techniques. The intent is to raise awareness of open questions in the field of AML and to call attention to trends that may shape risk and risk management practices in future.

NIST打算随着AML攻击和缓解措施的新发展而更新本报告。

本报告对PredAI和GenAI系统的常见攻击类别及其缓解措施进行了分类。本报告并非旨在对对抗性机器学习的全部现有文献进行详尽调查，截至2024年7月，arXiv.org上自2021年以来已有超过11,354篇相关参考文献。

本报告分为三个部分。

- 第2节讨论PredAI系统。第2.1节介绍PredAI系统的攻击分类法，该分类法定义了攻击者目的和目标的广泛类别，并确定了对手为实现相应目标必须利用的能力。还针对每种能力类型引入了具体的攻击类别。第2.2、2.3和2.4节分别讨论主要攻击类别：规避、投毒和隐私。各攻击类别章节中提供了针对每类攻击的相应缓解措施集。
- 第3节讨论GenAI系统。第3.1节介绍GenAI系统的攻击分类法，并定义与这些系统相关的攻击者目标和对抗者能力的广泛类别。针对每种能力类型介绍具体的攻击类别，并给出相关缓解措施。
- 第4节讨论该领域剩余的挑战，包括广泛使用的缓解技术的局限性。其目的是提高对AML领域开放问题的认识，并提请关注可能塑造未来风险与风险管理实践的趋势。

2. Predictive AI Taxonomy

2.1. Attack Classification

Figure 1 introduces a taxonomy of attacks in AML on PredAI systems, based on attacker goals and objectives, capabilities, and knowledge.

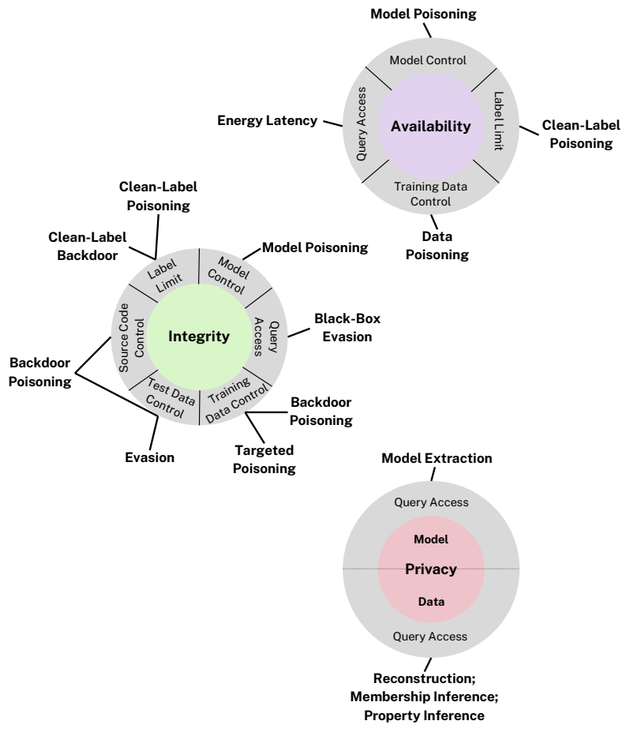


Figure 1. Taxonomy of attacks on PredAI systems

The attacker's objectives are shown as disjointed circles with the attacker's goal at the center of each circle: **availability** breakdown, **integrity** violation, and **privacy** compromise. The capabilities that an adversary must leverage to achieve their objectives are shown in the outer layer of the objective circles. Attack classes are shown as callouts connected to the capabilities required to mount each attack. Multiple attack classes that require the same capabilities to reach the same objective are shown in a single callout.

These attacks are classified according to the following dimensions: 1) learning method and stage of the learning process when the attack is mounted, 2) attacker goals and objectives,

2. 预测式 AI 分类体系

2.1. 攻击分类

图1介绍了PredAI系统上对抗性机器学习攻击的分类法，基于攻击者的目标和目的、能力和知识。

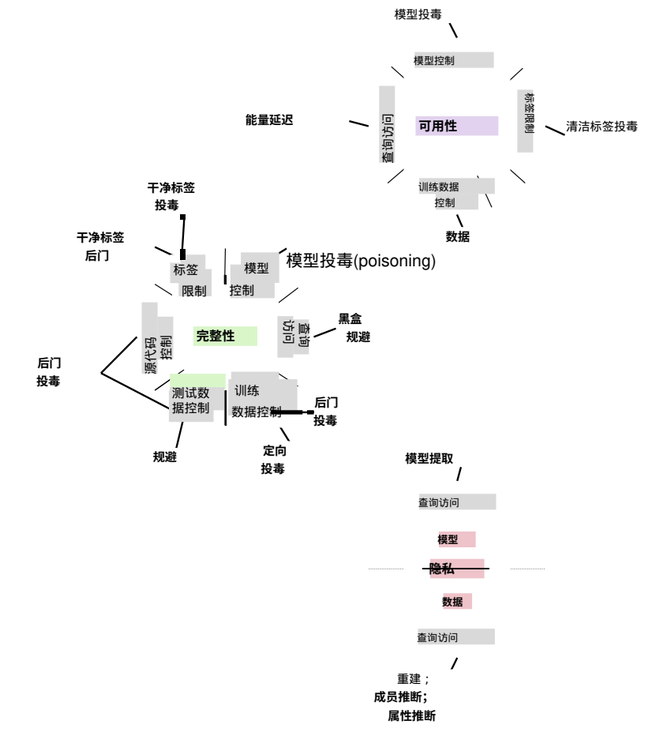


图 1. 针对 PredAI 系统的攻击分类体系

攻击者的目标以互不相交的圆圈表示，攻击者的目标位于每个圆圈的中心：**可用性**破坏、**完整性**违反和**隐私**泄露。对手为实现其目标必须利用的能力显示在目标圆圈的外层。攻击类别以标注形式显示，并与发动每种攻击所需的能力相连。需要相同能力达到相同目标的多个攻击类别显示在单个标注中。

这些攻击根据以下维度进行分类：1)学习方法及攻击实施时学习过程的阶段，2)攻击者的目标和目的，

3) attacker capabilities, and 4) attacker knowledge of the learning process. Several adversarial attack classification frameworks have been introduced in prior works [42, 358], and the goal here is to create a standard terminology for adversarial attacks on ML that unifies existing work.

2.1.1. Stages of Learning

Predictive machine learning involves a TRAINING STAGE in which a model is learned and a DEPLOYMENT STAGE in which the model is deployed on new, unlabeled data samples to generate predictions. In the case of SUPERVISED LEARNING, labeled training data is given as input to a training algorithm in the training stage, and the ML model is optimized to minimize a specific loss function. Validation and testing of the ML model is usually performed before the model is deployed in the real world. Common supervised learning techniques include CLASSIFICATION in which the predicted labels or *classes* are discrete and REGRESSION in which the predicted labels or *response variables* are continuous.

Other learning paradigms in the ML literature include UNSUPERVISED LEARNING, which trains models using unlabeled data at training time; SEMI-SUPERVISED LEARNING in which a small set of examples have labels, while the majority of samples are unlabeled; REINFORCEMENT LEARNING in which an agent interacts with an environment and learns an optimal policy to maximize its reward; FEDERATED LEARNING in which a set of clients jointly train an ML model by communicating with a server that performs an aggregation of model updates; and ENSEMBLE LEARNING, which is an approach that seeks better predictive performance by combining the predictions from multiple models.

Most PredAI models are DISCRIMINATIVE, i.e., learn only a decision boundary, such as LOGISTIC REGRESSION, SUPPORT VECTOR MACHINES, and CONVOLUTIONAL NEURAL NETWORKS. GenAI models may also be used in predictive tasks, such as sentiment analysis [125] .

AML literature predominantly considers adversarial attacks against AI systems that could occur at either the training stage or the deployment stage. During the training stage, the attacker might control part of the training data, their labels, the model parameters, or the code of ML algorithms, resulting in different types of poisoning attacks. During the deployment stage, the ML model is already trained, and the adversary could mount evasion attacks to create integrity violations and change the ML model's predictions, as well as privacy attacks to infer sensitive information about the training data or the ML model.

Training-time attacks. POISONING ATTACKS [40] occur during the ML training stage. In a DATA POISONING attack [40, 148], an adversary controls a subset of the training data by either inserting or modifying training samples. In a MODEL POISONING attack [222], the adversary controls the model and its parameters. Data poisoning attacks are applicable to all learning paradigms, while model poisoning attacks are most prevalent in federated learning [190], where clients send local model updates to the aggregating server, and in supply-chain attacks, where malicious code may be added to the model by suppliers of

3)攻击者能力，以及4)攻击者对学习过程的了解。先前工作中已引入了若干对抗性攻击分类框架[42,358]，此处的目标是建立一套统一的机器学习对抗性攻击标准术语，以整合现有工作。

2.1.1. 学习阶段

预测性机器学习包括一个训练阶段，在该阶段中学习一个模型，以及一个部署阶段，在该阶段中将模型部署到新的、未标记的数据样本上以生成预测。在监督学习的情况下，在训练阶段将带标签的训练数据作为输入提供给训练算法，并对机器学习模型进行优化以最小化特定的损失函数。机器学习模型的验证和测试通常在实际部署模型之前进行。常见的监督学习技术包括分类，其中预测的标签或类别是离散的，以及回归，其中预测的标签或响应变量是连续的。

机器学习文献中的其他学习范式包括无监督学习，其在训练时使用未标记数据训练模型；半监督学习，其中一小部分示例有标签，而大多数样本未标记；强化学习，其中智能体与环境交互并学习最优策略以最大化其奖励；联邦学习，其中一组客户端通过与执行模型更新聚合的服务器通信来联合训练一个机器学习模型；以及集成学习，这是一种通过组合多个模型的预测来寻求更好预测性能的方法。

大多数PredAI模型是判别式的，即仅学习决策边界，例如逻辑回归、支持向量机和卷积神经网络。GenAI模型也可用于预测任务，如情感分析[125]。

对抗性机器学习文献主要考虑针对人工智能系统的对抗攻击，这些攻击可能发生在训练阶段或部署阶段。在训练阶段，攻击者可能控制部分训练数据、其标签、模型参数或机器学习算法的代码，从而导致不同类型的投毒攻击。在部署阶段，机器学习模型已经训练完成，对手可以发动规避攻击以造成完整性违反并改变机器学习模型的预测，以及发动隐私攻击以推断关于训练数据或机器学习模型的敏感信息。

训练时攻击。投毒攻击[40]发生在机器学习训练阶段。在数据投毒攻击[40,148]中，对手通过插入或修改训练样本来控制训练数据的一个子集。在模型投毒攻击[222]中，对手控制模型及其参数。数据投毒攻击适用于所有学习范式，而模型投毒攻击在联邦学习[190]中最为普遍，其中客户端向聚合服务器发送本地模型更新，以及在供应链攻击中，供应商可能向模型添加恶意代码

model technology.

Deployment-time attacks. Other types of attacks can be mounted against deployed models. Evasion attacks modify testing samples to create ADVERSARIAL EXAMPLE [38, 144, 362], which are similar to the original sample (e.g., according to certain distance metrics) but alter the model predictions to the attacker's choices. Other attacks, such as availability attacks and privacy attacks including membership inference [342] and data reconstruction [110], can also be mounted by attackers with query access to a deployed ML model.

2.1.2. Attacker Goals and Objectives

The attacker's objectives are classified along three dimensions according to the three main types of security violations considered when analyzing the security of a system: availability breakdown, integrity violation, and privacy compromise. Figure 1 separates attacks into three disjointed circles according to their objective, and the attacker's objective is shown at the center of each circle.

Availability breakdown [NISTAML.01] [Back to Index]. An AVAILABILITY BREAKDOWN attack is a deliberate interference with a PredAI system to disrupt the ability of other users or processes to obtain timely and reliable access to its services. This attack type may be initiated at training or deployment time, although its impacts are typically experienced at deployment time. Availability attacks can be mounted via data poisoning, when the attacker controls a fraction of the training set; via model poisoning, when the attacker controls the model parameters; or as ENERGY-LATENCY ATTACK via query access. Data poisoning availability attacks have been proposed for SUPPORT VECTOR MACHINES [40], linear regression [179], and even neural networks [228, 260], while model poisoning attacks have been designed for neural networks [222] and federated learning [22].

- **Energy latency attacks [NISTAML.014] [Back to Index].** Recently, ENERGY-LATENCY ATTACK, a type of availability attacks that require only black-box access to the model, have been developed for neural networks across many different tasks in computer vision and natural language processing (NLP) [345].

Integrity violation [NISTAML.02] [Back to Index]. An INTEGRITY VIOLATION attack is a deliberate interference with a PredAI system to force it to misperform against its intended objectives and produce predictions that align with the adversary's objective. An attacker can cause an integrity violation by mounting an evasion attack at deployment time or a poisoning attack at training time. Evasion attacks require the modification of testing samples to create adversarial examples that are misclassified by the model while often remaining stealthy and imperceptible to humans [38, 144, 362]. Integrity attacks via poisoning can be classified as TARGETED POISONING ATTACK [137, 330], BACKDOOR POISONING ATTACK [148], and MODEL POISONING [22, 36, 123]. Targeted poisoning tries to violate the integrity of a few targeted samples and assumes that the attacker has training data control to insert the poisoned samples. Backdoor poisoning attacks require the generation of a BACKDOOR PATTERN,

模型技术。

部署时攻击。其他类型的攻击可以针对已部署的模型实施。规避攻击修改测试样本以创建对抗样本[38,144,362]，这些样本与原始样本相似(例如，根据某些距离度量)，但将模型预测改变为攻击者所选的结果。其他攻击，例如可用性攻击和隐私攻击(包括成员推断[342]和数据重构[110])，也可以由对已部署的ML模型具有查询访问权限的攻击者实施。

2.1.2. 攻击者的目标与目的

攻击者的目标根据分析系统安全性时考虑的三种主要安全违规类型沿三个维度进行分类：可用性破坏、完整性违规和隐私泄露。图1根据目标将攻击分为三个不相交的圆，攻击者的目标显示在每个圆的中心。

可用性破坏 [NISTAML.01] [返回索引] 。可用性破坏攻击是一种对PredAI系统的蓄意干扰，旨在破坏其他用户或进程及时、可靠地获取其服务的能力。此类攻击可能在训练时或部署时发起，但其影响通常在部署时才会显现。可用性攻击可以通过数据投毒实施，即攻击者控制一部分训练集；也可以通过模型投毒实施，即攻击者控制模型参数；或者作为能量-延迟攻击通过查询访问实施。针对支持向量机[40]、线性回归[179]乃至神经网络[228,260]的数据投毒可用性攻击已被提出，而模型投毒攻击则针对神经网络[222]和联邦学习[22]设计。

- **能量延迟攻击 [NISTAML.014] [返回索引]** 。最近，能量-延迟攻击，一种仅需对模型进行黑盒访问的可用性攻击，已被开发用于计算机视觉和自然语言处理(NLP)中许多不同任务的神经网络[345]。

完整性违反 [NISTAML.02] [返回索引] 。完整性违反攻击是一种对PredAI系统的蓄意干扰，迫使其偏离预期目标运行，并产生与攻击者目标一致的预测。攻击者可以通过在部署时发起规避攻击或在训练时发起投毒攻击来造成完整性违反。规避攻击需要修改测试样本来创建对抗样本，使模型对其产生误分类，同时通常保持隐蔽且人类难以察觉[38,144,362]。通过投毒实施的完整性攻击可分为定向投毒攻击[137,330]、后门投毒攻击[148]和模型投毒[22,36,123]。定向投毒试图违反少数定向样本的完整性，并假设攻击者拥有训练数据控制权以插入投毒样本。后门投毒攻击需要生成后门模式，

which is added to both the poisoned samples and the testing samples to cause misclassification. Backdoor attacks are the only attacks in the literature that require both training and testing data control. Model poisoning attacks could result in either targeted or backdoor attacks, and the attacker modifies model parameters to cause an integrity violation. They have been designed for centralized learning [222] and federated learning [22, 36].

Privacy compromise [NISTAML.03] [Back to Index]. A PRIVACY COMPROMISE attack causes the unintended leakage of restricted or proprietary information from a PredAI system, including details about a model's training data, weights, or architecture [100, 309]. While the term "confidentiality" is more widely used in taxonomies of traditional cybersecurity attacks, the AML field has tended to use the top-level term "privacy" to encompass both attacks against the confidentiality of a model (e.g., those that extract information about a model's weights or architecture) and those that cause violations of expected privacy properties of model outputs (e.g. by exposing model training data) [310]. DATA CONFIDENTIALITY during ML training can be achieved through secure computation methods based on cryptographic techniques [2, 253, 288, 385], which ensure that training data and model parameters remain protected during the training phase. However, even models trained using paradigms that enforce data confidentiality may be vulnerable to privacy attacks, in which adversaries interacting with a model can extract information about its training data or parameters. In this report, we focus on privacy compromises that can occur at deployment time, regardless of the training method used, or whether data confidentiality was maintained during training.

In privacy attacks, attackers might be interested in learning information about the training data (resulting in DATA PRIVACY ATTACKS) or the ML model (resulting in MODEL PRIVACY ATTACKS). The attacker could have different objectives for compromising the privacy of training data, such as DATA RECONSTRUCTION [110] (inferring the content or features of training data), MEMBERSHIP-INFERENCE ATTACK [162, 343] (inferring the presence of data in the training set), TRAINING DATA EXTRACTION [59, 63] (extracting training data from generative models), ATTRIBUTE INFERENCE ATTACKS [184, 409] (inferring sensitive attributes of training records) and PROPERTY INFERENCE [134] (inferring properties about the training data distribution). MODEL EXTRACTION is a model privacy attack in which attackers aim to extract information about the model [177].

2.1.3. Attacker Capabilities

AML attacks for PredAI systems can be taxonomized with respect to the capabilities that an attacker controls. An adversary might leverage six types of capabilities to achieve their objectives, as shown in the outer layer of the objective circles in Fig. 1:

- TRAINING DATA CONTROL: The attacker might take control of a subset of the training data by inserting or modifying training samples. This capability is used in data poisoning attacks (e.g., availability poisoning, targeted or backdoor poisoning).

该扰动被添加到投毒样本和测试样本中以导致误分类。后门攻击是文献中唯一需要同时控制训练和测试数据的攻击。模型投毒攻击可导致定向攻击或后门攻击，攻击者修改模型参数以造成完整性违规。它们已被设计用于集中式学习[222]和联邦学习[22,36]。

隐私泄露 [NISTAML.03] [返回索引] 。隐私泄露攻击导致PredAI系统意外泄露受限或专有信息，包括有关模型训练数据、权重或架构的细节[100,309]。虽然"机密性"一词在传统网络安全攻击分类中更为常用，但AML领域倾向于使用顶层术语"隐私"来涵盖针对模型机密性的攻击(例如提取模型权重或架构信息的攻击)以及导致模型输出违反预期隐私属性的攻击(例如暴露模型训练数据)[310]。ML训练期间的数据机密性可以通过基于密码学技术的安全计算方法实现[2,253,288,385]，这些方法确保训练数据和模型参数在训练阶段受到保护。然而，即使使用强制数据机密性的范式训练的模型也可能容易受到隐私攻击，攻击者通过与模型交互可以提取有关其训练数据或参数的信息。在本报告中，我们关注部署时可能发生的隐私泄露，无论使用何种训练方法，也无论训练期间是否保持了数据机密性。

在隐私攻击中，攻击者可能对了解训练数据的信息感兴趣(导致数据隐私攻击)或对ML模型的信息感兴趣(导致模型隐私攻击)。攻击者可能具有不同的目标来破坏训练数据的隐私，例如数据重构[110](推断训练数据的内容或特征)、成员推断攻击[162,343](推断训练集中数据的存在)、训练数据提取[59,63](从生成模型中提取训练数据)、属性推断攻击[184,409](推断训练记录的敏感属性)和属性推断[134](推断有关训练数据分布的属性)。模型萃取是一种模型隐私攻击，攻击者旨在提取有关模型的信息[177]。

2.1.3. 攻击者能力

PredAI系统的对抗性机器学习攻击可以根据攻击者所控制的能力进行分类。对手可能利用六种能力来实现其目标，如图1中目标圆的外层所示：

- 训练数据控制：攻击者可能通过插入或修改训练样本来控制训练数据的子集。此能力用于数据投毒攻击(例如，可用性投毒、定向投毒或后门投毒)。

- MODEL CONTROL: The attacker might take control of the model parameters by either generating a Trojan trigger and inserting it in the model or by sending malicious local model updates in federated learning.
- TESTING DATA CONTROL: The attacker might add perturbations to testing samples at model deployment time, as performed in evasion attacks to generate adversarial examples or in backdoor poisoning attacks.
- LABEL LIMIT: This capability is relevant to restrict adversarial control over the labels of training samples in supervised learning. Clean-label poisoning attacks assume that the attacker does not control the label of the poisoned samples, while regular poisoning attacks assume label control over the poisoned samples.
- SOURCE CODE CONTROL: The attacker might modify the source code of the ML algorithm, such as the random number generator or any third-party libraries, which are often open source.
- QUERY ACCESS: The attacker might submit queries to the model and receive predictions (i.e., labels or model confidences), such as when interacting with an AI system hosted by a cloud provider as a machine learning as a service (MLaaS) offering. This capability is used by black-box evasion attacks, ENERGY-LATENCY ATTACK, and all privacy attacks that do not require knowledge of the model's training data, architecture, or parameters.

Even if an attacker does not have the ability to modify training/testing data, source code, or model parameters, access to these may still be crucial for mounting stronger white-box attacks that require knowledge of the ML system. See Sec. 2.1.4 for more details on attacker knowledge, and detailed definitions of white-box and black-box attacks.

Figure 1 connects each attack class with the capabilities required to mount the attack. For example, backdoor attacks that cause integrity violations require control of the training and testing data to insert the backdoor pattern. Backdoor attacks can also be mounted via source code control, particularly when training is outsourced to a more powerful entity. Clean-label backdoor attacks do not allow label control on the poisoned samples in addition to the capabilities needed for backdoor attacks.

2.1.4. Attacker Knowledge

Another dimension of attack classification is how much knowledge the attacker has about the ML system. There are three main types of attacks:

White-box attacks. These assume that the attacker operates with *full* knowledge about the ML system, including the training data, model architecture, and model hyperparameters. While these attacks operate under very strong assumptions, the main reason for analyzing them is to test the vulnerability of a system against worst-case adversaries and

- 模型控制：攻击者可能通过生成木马触发器并将其插入模型，或在联邦学习中发送恶意本地模型更新来控制模型参数。
- 测试数据控制：攻击者可能在模型部署时向测试样本添加扰动，如在规避攻击中生成对抗样本或在后门投毒攻击中所做的那样。
- 标签限制：此能力与监督学习中限制对训练样本标签的对抗性控制相关。干净标签投毒攻击假设攻击者不控制投毒样本的标签，而常规投毒攻击假设攻击者控制投毒样本的标签。
- 源代码控制：攻击者可能修改机器学习算法的源代码
  算法，例如随机数生成器或任何第三方库，这些通常是开源的。
- 查询访问：攻击者可能向模型提交查询并接收预测
  结果(即标签或模型置信度)，例如在与云提供商托管的、作为机器学习即服务(MLaaS)产品的人工智能系统交互时。这种能力被黑盒规避攻击、ENERGY-LATENCY ATTACK以及所有不需要了解模型训练数据、架构或参数的隐私攻击所使用。

即使攻击者没有能力修改训练/测试数据、源代码或模型参数，访问这些内容对于发动需要了解机器学习系统的更强的白盒攻击可能仍然至关重要。有关攻击者知识的更多细节以及白盒和黑盒攻击的详细定义，请参见第2.1.4节。

图1将每个攻击类别与实施该攻击所需的能力联系起来。例如，造成完整性违反的后门攻击需要控制训练和测试数据以插入后门模式。后门攻击也可以通过源代码控制来实施，特别是当训练外包给更强大的实体时。干净标签后门攻击除了需要后门攻击所需的能力外，还不允许对投毒样本进行标签控制。

2.1.4. 攻击者知识

攻击分类的另一个维度是攻击者对机器学习系统有多少了解。主要有三种类型的攻击：

白盒攻击。这些攻击假设攻击者在完全了解机器学习系统的情况下进行操作，包括训练数据、模型架构和模型超参数。虽然这些攻击在非常强的假设下运行，但分析它们的主要原因是测试系统对最坏情况对手的脆弱性，以及

to evaluate potential mitigations. This definition is more general and encompasses the notion of adaptive attacks in which knowledge of the mitigations applied to the model or the system is explicitly tracked.

Black-box attacks. These attacks assume that the attacker operates with minimal, and sometimes no knowledge at all about the ML system. An adversary might have query access to the model, but they have no other information about how the model is trained. These attacks are the most practical since they assume that the attacker has no knowledge of the AI system and utilizes system interfaces readily available for normal use.

Gray-box attacks. There are a range of gray-box attacks that capture adversarial knowledge between black-box and white-box attacks. Suciu et al. [358] introduced a framework to classify gray-box attacks. An attacker might know the model architecture but not its parameters, or the attacker might know the model and its parameters but not the training data. Other common assumptions for gray-box attacks are that the attacker has access to data distributed identically to the training data and knows the feature representation. The latter assumption is important for applications in which feature extraction is used before training an ML model, such as cybersecurity, finance, and healthcare.

2.1.5. Data Modality

Until recently, most attacks and defenses in adversarial machine learning have operated under a single modality, but a new trend in the field is to use multimodal data. The taxonomy of attacks defined in Fig. 1 is independent of the modality of the data in specific applications.

The most common data modalities in the AML literature include:

- **Image:** Adversarial examples of image data [144, 362] have the advantage of a continuous domain, and gradient-based methods can be applied directly for optimization. Backdoor poisoning attacks were first invented for images [148], and many privacy attacks are run on image datasets (e.g., [342]). The image modality includes other types of imaging (e.g., LIDAR, SAR, IR, hyperspectral).
- **Text:** Text is a popular modality, and all classes of attacks have been proposed for text models, including evasion [150], poisoning [82, 213], and privacy [426].
- **Audio:** Audio systems and text generated from audio signals have also been attacked [66].
- **Video:** Video comprehension models have shown increasing capabilities in vision and language tasks [428], but such models are also vulnerable to attacks [402].
- **Cybersecurity**[2]**:** The first poisoning attacks were discovered in cybersecurity for worm

[2]Cybersecurity data may not include a single modality but rather multiple modalities, such as network-level, host-level, or program-level data.

评估潜在的缓解措施。这个定义更为通用，涵盖了自适应攻击的概念，其中对应用于模型或系统的缓解措施的知识被明确追踪。

黑盒攻击。这些攻击假设攻击者在极少、有时甚至完全不了解机器学习系统的情况下进行操作。对手可能对模型有查询访问权限，但他们没有关于模型如何训练的其他信息。这些攻击是最实用的，因为它们假设攻击者不了解人工智能系统，并利用正常使用中随时可用的系统接口。

灰盒攻击。存在一系列灰盒攻击，它们捕捉了黑盒攻击和白盒攻击之间的对抗知识。Suciu等人[358]提出了一个框架来对灰盒攻击进行分类。攻击者可能知道模型架构但不知道其参数，或者攻击者可能知道模型及其参数但不知道训练数据。灰盒攻击的其他常见假设是攻击者可以访问与训练数据同分布的数据并知道特征表示。后一个假设对于在训练ML模型之前使用特征提取的应用很重要，例如网络安全、金融和医疗保健。

2.1.5. 数据模态

直到最近，对抗性机器学习中的大多数攻击和防御都是在单一模态下运行的，但该领域的一个新趋势是使用多模态数据。图1中定义的攻击分类法独立于特定应用中数据的模态。

AML 文献中最常见的数据模态包括：

- **图像：**图像数据的对抗样本[144,362]具有连续域的优势，基于梯度的方法可以直接应用于优化。后门投毒攻击最初是为图像发明的[148]，许多隐私攻击是在图像数据集上运行的(例如[342])。图像模态包括其他类型的成像(例如LIDAR、SAR、IR、高光谱)。
- **文本：**文本是一种流行的模态，所有类别的攻击都已被提出用于文本模型，包括规避[150]、投毒[82,213]和隐私[426]。
- **音频：**音频系统和从音频信号生成的文本也一直钉住了 [66]。
- **视频：**视频理解模型在视觉和语言任务中已展现出越来越强的能力[428]，但此类模型也容易受到攻击[402]。
- **网络安全**[2]**:**最早的投毒攻击是在网络安全领域针对蠕虫

2 网络安全数据可能不包含单一模态，而是多种模态，例如网络级、主机级或程序级数据。

signature generation (2006) [291] and spam email classification (2008) [269]. Since then, poisoning attacks have been shown for malware classification, malicious PDF detection, and Android malicious app classification [329]. Evasion attacks against similar data modalities have been proposed as well: malware classification [103, 357], PDF malware classification [352, 414], Android malicious app detection [295], and network instrusion detection [93]. Poisoning unsupervised learning models has been shown for clustering used in malware classification [41] and network traffic anomaly detection [315].

Anomaly detection based on data-centric approaches allows for automated feature learning through ML algorithms. However, the application of ML to such problems comes with specific challenges related to the need for very low false negative and low false positive rates (e.g., the ability to catch zero-day attacks). This challenge is compounded by the fact that trying to accommodate all of these together makes ML models susceptible to adversarial attacks [198, 301, 446].

- **Tabular data:** There have been numerous attacks against ML models working on tabular data, such as poisoning availability attacks against healthcare and business applications [179], privacy attacks against healthcare data [422], and evasion attacks against financial applications [141].

Recently, the use of ML models trained on multimodal data has gained traction, particularly the combination of image and text data modalities. Several papers have shown that multimodal models may provide some resilience against attacks [417], but other papers show that multimodal models themselves could be vulnerable to attacks mounted on all modalities at the same time [77, 333, 415] (see Sec. 4.2.3).

An open challenge is to test and characterize the resilience of a variety of multimodal ML models against evasion, poisoning, and privacy attacks.

签名生成(2006)[291]和垃圾邮件分类(2008)[269]。此后，投毒攻击已被证明可用于恶意软件分类、恶意PDF检测和Android恶意应用分类[329]。针对类似数据模态的规避攻击也已被提出：恶意软件分类[103,357]、PDF恶意软件分类[352,414]、Android恶意应用检测[295]和网络入侵检测[93]。针对无监督学习模型的投毒已被证明可用于恶意软件分类[41]和网络流量异常检测[315]中使用的聚类。

基于数据中心方法的异常检测允许通过机器学习算法进行自动特征学习。然而，将机器学习应用于此类问题会带来特定挑战，这与需要极低的假阴性率和低假阳性率有关(例如，捕获零日攻击的能力)。这一挑战因以下事实而加剧：试图同时满足所有这些要求会使机器学习模型容易受到对抗攻击[198,301,446]。

- **表格数据：**已有大量针对处理表格数据的机器学习模型的攻击，例如针对医疗和商业应用的投毒可用性攻击[179]、针对医疗数据的隐私攻击[422]，以及针对金融应用的规避攻击[141]。

最近，使用在多模态数据上训练的机器学习模型的做法获得了关注，特别是图像和文本数据模态的组合。几篇论文表明，多模态模型可能提供一定的抗攻击能力[417]，但其他论文表明，多模态模型本身可能容易受到同时针对所有模态发动的攻击[77,333,415](参见第4.2.3节)。

一个开放性挑战是测试和表征各种多模态机器学习模型对规避、投毒和隐私攻击的韧性。

2.2. Evasion Attacks and Mitigations

[**NISTAML.022**] [Back to Index]

The discovery of evasion attacks against ML models has led to significant growth in AML research over the last decade. In an evasion attack, the adversary's goal is to generate adversarial examples: samples whose classification can be changed to an arbitrary class of the attacker's choice – often with only minimal perturbation [362]. For example, in the context of image classification, the perturbation of the original sample might be small so that a human cannot observe the transformation of the input; while the ML model can be tricked to classify the adversarial example in the target class selected by the attacker, humans still recognize it as part of the original class.

Early known instances of evasion attacks date back to 1988 with the work of Kearns and Li [192] and 2004 when Dalvi et al. [98] and Lowd and Meek [226] demonstrated the existence of adversarial examples for linear classifiers used in spam filters. Later, Szedegy et al. [362] showed that deep neural networks used for image classification could be easily manipulated through adversarial examples. In 2013, Szedegy et al. [362] and Biggio et al. [38] independently discovered an effective method for generating adversarial examples against linear models and neural networks by applying gradient optimization to an adversarial objective function. Both of these techniques require white-box access to the model and were improved by subsequent methods that generated adversarial examples with even smaller perturbations [20, 65, 232].

Adversarial examples are also applicable in more realistic black-box settings in which attackers only obtain query access capabilities to the trained model. Even in the more challenging black-box setting in which attackers obtain the model's predicted labels or confidence scores, deep neural networks are still vulnerable to adversarial examples. Methods for creating adversarial examples in black-box settings include zeroth-order optimization [80], discrete optimization [254], Bayesian optimization [344], and *transferability*, which involves the white-box generation of adversarial examples on a different model before transferring them to the target model [282, 283, 377]. While cybersecurity and image classifications were the first application domains to showcase evasion attacks, ML technology in many other application domains has come under scrutiny, including speech recognition [66], natural language processing [185], and video classification [215, 401].

Mitigating adversarial examples is a well-known challenge in the community and deserves additional research and investigation. The field has a history of publishing defenses evaluated under relatively weak adversarial models that are subsequently broken by more powerful attacks. Mitigations need to be evaluated against strong adaptive attacks, and guidelines for the rigorous evaluation of newly proposed mitigation techniques have been established [97, 375]. The most promising directions for mitigating the critical threat of evasion attacks are adversarial training [144, 232] (iteratively generating and inserting adversarial examples with their correct labels at training time); certified techniques, such as

2.2. 规避攻击与缓解

[**NISTAML.022**][返回索引]

针对ML模型的规避攻击的发现导致了过去十年中AML研究的显著增长。在规避攻击中，对手的目标是生成对抗样本：其分类可以被改变为攻击者选择的任意类别的样本--通常只需最小的扰动[362]。例如，在图像分类的背景下，原始样本的扰动可能很小，以至于人类无法观察到输入的变换；而ML模型可能被欺骗将对抗样本分类为攻击者选择的目标类别，人类仍然将其识别为原始类别的一部分。

早期已知的规避攻击实例可以追溯到1988年Kearns和Li的工作[192]以及2004年Dalvi等人[98]和Lowd和Meek[226]证明了用于垃圾邮件过滤器的线性分类器存在对抗样本。后来，Szedegy等人[362]表明用于图像分类的深度神经网络可以通过对抗样本轻易操纵。2013年，Szedegy等人[362]和Biggio等人[38]独立发现了一种有效的方法，通过对对抗目标函数应用梯度优化来生成针对线性模型和神经网络的对抗样本。这两种技术都需要对模型的白盒访问权限，并被后续方法改进，这些方法生成了具有更小扰动的对抗样本[20,65,232]。

对抗样本同样适用于更现实的黑盒设置，其中攻击者仅获得对已训练模型的查询访问能力。即使在更具挑战性的黑盒设置中--攻击者获得模型的预测标签或置信度分数--深度神经网络仍然容易受到对抗样本的影响。在黑盒设置中创建对抗样本的方法包括零阶优化[80]、离散优化[254]、贝叶斯优化[344]以及迁移性，后者涉及在不同模型上以白盒方式生成对抗样本，然后将其迁移到目标模型[282,283,377]。虽然网络安全和图像分类是最早展示规避攻击的应用领域，但许多其他应用领域的ML技术也受到了审视，包括语音识别[66]、自然语言处理[185]和视频分类[215,401]。

缓解对抗样本是社区中众所周知的挑战，值得进一步研究和调查。该领域历史上发表过在相对较弱的对抗模型下评估的防御措施，随后被更强大的攻击所攻破。缓解措施需要针对强自适应攻击进行评估，并且已经建立了严格评估新提出的缓解技术的指南[97,375]。缓解规避攻击这一关键威胁最有前景的方向是对抗训练[144,232](在训练时迭代生成并插入带有正确标签的对抗样本)；认证技术，如

randomized smoothing [94] (evaluating ML prediction under noise); and formal verification techniques [136, 191] (applying formal method techniques to verify the model's output). Nevertheless, these methods have different limitations, such as decreased accuracy for adversarial training and randomized smoothing and computational complexity for formal methods. There is an inherent trade-off between robustness and accuracy [374, 379, 433]. Similarly, there are trade-offs between a model's robustness and fairness guarantees [71].

2.2.1. White-Box Evasion Attacks

In the white-box threat model, the attacker has full knowledge of the model architecture and parameters, as discussed in Section 2.1.4. The main challenge for creating adversarial examples in this setting is to find a perturbation added to a testing sample that changes its classification label, often with constraints on properties such as the perceptibility or size of the perturbation. In the white-box threat model, it is common to craft adversarial examples by solving an optimization problem written from the attacker's perspective, which specifies the objective function for the optimization (such as changing the target label to a certain class), as well as a distance metric to measure the similarity between the testing sample and the adversarial example.

Optimization-based methods. Szedegy et al. [362] and Biggio et al. [38] independently proposed the use of optimization techniques to generate adversarial examples. In their threat models, the adversary is allowed to inspect the entirety of the ML model and compute gradients relative to the model's loss function. These attacks can be targeted (i.e., the adversarial example's class is selected by the attacker) or untargeted (i.e., the adversarial examples are misclassified to any other incorrect class).

Szedegy et al. [362] coined the widely used term *adversarial examples*. They considered an objective that minimized the ℓ_2 norm of the perturbation subject to the model prediction changing to the target class. The optimization is solved using the limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) method. Biggio et al. [38] considered the setting of a binary classifier with malicious and benign classes with a continuous and differentiable discriminant function. The objective of the optimization is to minimize the discriminant function in order to generate adversarial examples of maximum confidence.

While Biggio et al. [38] applied their method to linear classifiers, kernel SVM, and multilayer perceptrons, Szedegy et al. [362] showed the existence of adversarial examples on deep learning models used for image classification. Goodfellow et al. [144] introduced an efficient method for generating adversarial examples for deep learning: the Fast Gradient Sign Method (FGSM), which performs a single iteration of gradient descent for solving the optimization. This method has been extended to an iterative FGSM attack by Kurakin et al. [200].

Subsequent works have proposed new objectives and methods for optimizing the generation of adversarial examples with the goals of minimizing the perturbations and supporting

随机平滑[94](在噪声下评估ML预测)；以及形式化验证技术[136,191](应用形式化方法技术来验证模型的输出)。然而，这些方法具有不同的局限性，例如对抗训练和随机平滑的准确率下降，以及形式化方法的计算复杂性。鲁棒性与准确率之间存在固有的权衡[374,379,433]。同样，模型的鲁棒性与公平性保证之间也存在权衡[71]。

2.2.1. 白盒规避攻击

在白盒威胁模型中，攻击者完全了解模型架构和参数，如第2.1.4节所述。在此设置中创建对抗样本的主要挑战是找到添加到测试样本中的扰动，使其分类标签发生变化，通常对扰动的可感知性或大小等属性有约束。在白盒威胁模型中，通常通过求解从攻击者角度编写的优化问题来构造对抗样本，该问题指定了优化的目标函数(例如将目标标签更改为某个类别)，以及用于度量测试样本与对抗样本之间相似度的距离度量。

基于优化的方法。 Szegedy等人[362]和Biggio等人[38]独立提出了使用优化技术来生成对抗样本。在他们的威胁模型中，对手被允许检查ML模型的全部内容并计算相对于模型损失函数的梯度。这些攻击可以是有目标的(即对抗样本的类别由攻击者选择)或无目标的(即对抗样本被误分类为任何其他错误类别)。

Szedegy 等人[362]创造了广泛使用的术语对抗样本。他们考虑了一个目标，即在模型预测改变为目标类别的约束下，最小化扰动的 ` 2 范数。该优化使用有限内存 Broy-den-Fletcher-Goldfarb-Shanno(L-BFGS)方法求解。Biggio 等人[38]考虑了具有恶意类和良性类、且判别函数连续可微的二分类器设置。优化的目标是最小化判别函数，以生成具有最大置信度的对抗样本。

Biggio 等人[38]将其方法应用于线性分类器、核 SVM 和多层感知机，而 Szedegy 等人[362]证明了用于图像分类的深度学习模型上存在对抗样本。Goodfellow 等人[144]引入了一种为深度学习生成对抗样本的高效方法：快速梯度符号法(FGSM)，该方法执行一次梯度下降迭代来求解优化。该方法被 Kurakin 等人[200]扩展为迭代 FGSM 攻击。

后续工作提出了新的目标和方法来优化对抗样本的生成，目标是最小化扰动并支持

multiple distance metrics. Some notable attacks include:

- DeepFool is an untargeted evasion attack for ℓ_2 norms, which uses a linear approximation of the neural network to construct the adversarial examples [257].
- The Carlini-Wagner attack uses multiple objectives that minimize the loss or logits on the target class and the distance between the adversarial example and original sample. The attack is optimized via the penalty method [65] and considers three distance metrics to measure the perturbations of adversarial examples: ℓ_0, ℓ_2, and ℓ_∞. The attack has been effective against the defensive distillation defense [284].
- The Projected Gradient Descent (PGD) attack [232] minimizes the loss function and projects the adversarial examples to the space of allowed perturbations at each iteration of gradient descent. PGD can be applied to the ℓ_2 and ℓ_∞ distance metrics for measuring the perturbation of adversarial examples.

Universal evasion attacks. Moosavi-Dezfooli et al. [256] showed how to construct small universal perturbations (with respect to some norm) that can be added to most images and induce a misclassification. Their technique relies on successive optimization of the universal perturbation using a set of points sampled from the data distribution. This is a form of FUNCTIONAL ATTACK. An interesting observation is that the universal perturbations generalize across deep network architectures, suggesting similarity in the decision boundaries trained by different models for the same task.

Physically realizable attacks. These are attacks against ML systems that can be implemented feasibly in the physical world [21, 200, 227]. One of the first instances was the attack on facial recognition systems by Sharif et al. [332]. The attack can be realized by printing a pair of eyeglass frames, which misleads facial recognition systems to either evade detection or impersonate another individual. Eykholt et al. [122] proposed an attack to generate robust perturbations under different conditions, resulting in adversarial examples that can evade vision classifiers in various physical environments. The attack is applied to evade a road sign detection classifier by physically applying black and white stickers to the road signs. The ShapeShifter [81] attack was designed to evade object detectors, which is a more challenging problem than attacking image classifiers since the attacker needs to evade the classification in multiple bounding boxes with different scales. This attack also requires the perturbation to be robust enough to survive real-world distortions due to different viewing distances, angles, lighting conditions, and camera limitations.

Other data modalities. In computer vision applications, adversarial examples are often designed to be imperceptible to humans. Therefore, the perturbations introduced by attackers need to be so small that a human correctly recognizes the images, while the ML classifier is tricked into changing its prediction. Alternatively, there may be a trigger object in the image that is still imperceptible or innocuous to humans but causes the model to misclassify. The concept of adversarial examples has been extended to other domains, such as audio, video, NLP, and cybersecurity. In some of these settings, there are additional

多种距离度量。一些值得注意的攻击包括：

- DeepFool 是一种针对 ` 2 范数的无目标规避(evasion)攻击，其使用线性近似神经网络的优化以构建对抗样本。
- Carlini-Wagner 攻击使用多个目标，最小化目标类上的损失或 logits 以及对抗样本与原始样本之间的距离。该攻击通过惩罚方法[65]进行优化，并考虑三种距离度量来衡量对抗样本的扰动：` 0 、 ` 2 、和 ` ∞。该攻击对防御性蒸馏防御[284]有效。
- 投影梯度下降(PGD)攻击[232]最小化损失函数，并在梯度下降的每次迭代中将对抗样本投影到允许扰动的空间。PGD 可应用于 ` 2 和` ∞ 距离度量，用于衡量对抗样本的扰动。

通用规避攻击。Moosavi-Dezfooli等人[256]展示了如何构造小型通用扰动(相对于某种范数)，该扰动可添加到大多数图像中并诱导误分类。他们的技术依赖于使用从数据分布中采样的一组点对通用扰动进行逐次优化。这是一种功能性攻击形式。一个有趣的观察是，通用扰动可以跨深度网络架构泛化，这表明不同模型针对同一任务训练出的决策边界具有相似性。

物理可实现攻击。这些是针对 ML 系统的攻击，可以在物理世界中切实实施[21,200,227]。最早的实例之一是 Sharif 等人[332]对人脸识别系统的攻击。该攻击可以通过打印一副眼镜框来实现，从而误导人脸识别系统要么规避检测，要么冒充他人。Eykholt 等人[122]提出了一种在不同条件下生成稳健扰动的攻击，产生的对抗样本可以在各种物理环境中规避视觉分类器。该攻击通过物理地在道路标志上粘贴黑白贴纸，来规避道路标志检测分类器。ShapeShifter[81]攻击旨在规避目标检测器，这比攻击图像分类器更具挑战性，因为攻击者需要在多个不同尺度的边界框中规避分类。该攻击还要求扰动足够稳健，以经受住由于不同观察距离、角度、光照条件和相机限制造成的真实世界失真。

其他数据模态。在计算机视觉应用中，对抗样本通常被设计为人类不可感知。因此，攻击者引入的扰动需要小到人类能够正确识别图像，同时 ML 分类器被欺骗而改变其预测。或者，图像中可能存在一个对人类仍然不可感知或无害的触发对象，但会导致模型误分类。对抗样本的概念已扩展到其他领域，如音频、视频、NLP 和网络安全。在其中一些设置中，存在额外的

constraints that need to be respected by adversarial examples, such as text semantics in NLP and the application constraints in cybersecurity. Several representative works include:

- **Audio:** Carlini and Wagner [66] showed a targeted attack on models that generate text from speech. They can generate an audio waveform that is very similar to an existing one but that can be transcribed to any text of the attacker's choice.
- **Video:** Adversarial evasion attacks against video classification models can be split into sparse attacks that perturb a small number of video frames [401] and dense attacks that perturb all of the frames in a video [215]. The goal of the attacker is to change the classification label of the video.
- **Text:** Jia and Liang [185] developed a methodology for generating adversarial text examples. This pioneering work was followed by many advances in developing adversarial attacks on natural language processing (NLP) models (see a comprehensive survey on the topic [438]). La Malfa and Kwiatkowska [202] proposed a method for formalizing perturbation definitions in NLP by introducing the concept of semantic robustness. The main challenges in NLP are that the domain is discrete rather than continuous (e.g., image, audio, and video classification), and adversarial examples need to respect text semantics. These challenges are illustrated by the recent ASCII-art attack [186] against chatbots. An ASCII-art illustration of a forbidden term tricks the chatbot into providing the harmful information even when the chatbot correctly censors the plain English word. The semantic distance between the two prompts is precisely zero, and both of them should have been treated the same.
- **Cybersecurity:** In cybersecurity applications, adversarial examples must respect the constraints imposed by the application semantics and feature representation of cyber data, such as network traffic or program binaries. FENCE is a general framework for crafting white-box evasion attacks using gradient optimization in discrete domains and supports a range of linear and statistical feature dependencies [88]. FENCE has been applied to two network security applications: malicious domain detection and malicious network traffic classification. Sheatsley et al. [334] proposed a method that learns the constraints in feature space using formal logic and crafts adversarial examples by projecting them onto a constraint-compliant space. They applied the technique to network intrusion detection and phishing classifiers. Both papers observed that attacks from continuous domains cannot be readily applied in constrained environments, as they result in infeasible adversarial examples. Pierazzi et al. [295] discussed the difficulty of mounting feasible evasion attacks in cybersecurity due to constraints in feature space and the challenge of mapping attacks from feature space to problem space. They formalized evasion attacks in problem space and constructed feasible adversarial examples for Android malware.

对抗样本需要遵守的约束，例如自然语言处理中的文本语义和网络安全中的应用约束。几项代表性工作包括：

- **音频：**Carlini和Wagner[66]展示了对从语音生成文本的模型的有目标攻击。他们可以生成与现有波形非常相似的音频波形，但可以转录为攻击者选择的任何文本。
- **视频：**针对视频分类模型的对抗规避攻击可分为稀疏攻击(扰动少量视频帧[401])和密集攻击(扰动视频中的所有帧[215])。攻击者的目标是改变视频的分类标签。
- **文本：** Jia 和 Liang[185] 提出了一种生成对抗文本样本的方法。这项开创性工作之后，在开发针对自然语言处理(NLP)模型的对抗攻击方面出现了许多进展(参见关于该主题的全面综述[438])。La Malfa 和 Kwiatkowska[202] 通过引入语义鲁棒性的概念，提出了一种形式化 NLP 中扰动定义的方法。NLP 中的主要挑战在于该领域是离散的而非连续的(例如图像、音频和视频分类)，并且对抗样本需要尊重文本语义。这些挑战由近期针对聊天机器人的 ASCII 艺术攻击[186] 所说明。对禁用词的 ASCII 艺术图示会诱骗聊天机器人提供有害信息，即使聊天机器人正确地审查了纯英文单词。两个提示之间的语义距离恰好为零，二者本应被同等对待。
- **网络安全：**在网络安全应用中，对抗样本必须遵守由网络数据(如网络流量或程序二进制文件)的应用语义和特征表示所施加的约束。FENCE 是一个通用框架，用于在离散域中使用梯度优化构造白盒规避攻击，并支持一系列线性和统计特征依赖关系[88]。FENCE 已应用于两个网络安全应用：恶意域名检测和恶意网络流量分类。Sheatsley 等人[334] 提出了一种方法，使用形式逻辑学习特征空间中的约束，并通过将对抗样本投影到符合约束的空间来构造对抗样本。他们将这项技术应用于网络入侵检测和钓鱼分类器。两篇论文都观察到，来自连续域的攻击无法轻易应用于受约束环境，因为它们会导致不可行的对抗样本。Pierazzi 等人[295] 讨论了由于特征空间中的约束而在网络安全中实施可行规避攻击的困难，以及将攻击从特征空间映射到问题空间的挑战。他们在问题空间中形式化了规避攻击，并为 Android 恶意软件构造了可行的对抗样本。

2.2.2. Black-Box Evasion Attacks

[**NISTAML.025**] [Back to Index]

Black-box evasion attacks are designed under a realistic adversarial model in which the attacker has no prior knowledge of the model architecture or training data. Instead, the adversary can interact with a trained ML model by querying it on various data samples and obtaining the model's predictions. Similar APIs are provided by MLaaS offered by public cloud providers, in which users can obtain the model's predictions on selected queries without information about how the model was trained. There are two main classes of black-box evasion attacks in the literature:

- **Score-based attacks:** In this setting, attackers obtain the model's confidence scores or logits and can use various optimization techniques to create the adversarial examples. A popular method is zeroth-order optimization, which estimates the model's gradients without explicitly computing derivatives [80, 173]. Other optimization techniques include discrete optimization [254], natural evolution strategies [172], and random walks [262].
- **Decision-based attacks:** In this more restrictive setting, attackers only obtain the final predicted labels of the model. The first method for generating evasion attacks was the Boundary Attack based on random walks along the decision boundary and rejection sampling [47], which was extended with an improved gradient estimation to reduce the number of queries in the HopSkipJumpAttack [79]. More recently, several optimization methods search for the direction of the nearest decision boundary (e.g., the OPT attack [86]), use sign SGD instead of binary searches (e.g., the Sign-OPT attack [87]), or use Bayesian optimization [344].

The primary challenge in creating adversarial examples in black-box settings is reducing the number of queries to the ML models. Recent techniques can successfully evade the ML classifiers with a relatively small number of queries, typically less than 1000 [344].

2.2.3. Transferability of Attacks

Another method for generating adversarial attacks under restrictive threat models involves transferring an attack crafted on a different ML model. Typically, an attacker trains a substitute ML model, generates white-box adversarial attacks on the substitute model, and transfers the attacks to the target model. Various methods differ in how the substitute models are trained. For example, Papernot et al. [282, 283] train the substitute model with score-based queries to the target model, while several papers train an ensemble of models without explicitly querying the target model [218, 377, 397].

Attack transferability is an intriguing phenomenon, and existing literature attempts to un-

2.2.2. 黑盒规避攻击

[**NISTAML.025**][返回索引]

黑盒规避攻击是在现实的对抗模型下设计的，其中攻击者事先不知道模型架构或训练数据。相反，对手可以通过在各种数据样本上查询训练好的 ML 模型并获取模型预测来与其交互。公共云提供商提供的 MLaaS 也提供类似 API，用户可以在所选查询上获得模型预测，而无需了解模型是如何训练的。文献中有两大类黑盒规避攻击：

- **基于分数的攻击：**在此设置中，攻击者获得模型的置信度分数或logits，并可以使用各种优化技术来创建对抗样本。一种流行的方法是零阶优化，它在不显式计算导数的情况下估计模型的梯度[80,173]。其他优化技术包括离散优化[254]、自然进化策略[172]和随机游走[262]。
- **基于决策的攻击：**在这种更受限的设置中，攻击者只能获得模型的最终预测标签。生成规避攻击的第一种方法是基于沿决策边界随机游走和拒绝采样的边界攻击[47]，该方法通过改进的梯度估计进行了扩展，以减少HopSkipJumpAttack[79] 中的查询次数。最近，若干优化方法搜索最近决策边界的方向(例如 OPT 攻击[86])，使用符号 SGD 代替二分搜索(例如 Sign-OPT 攻击[87])，或使用贝叶斯优化[344]。

在黑盒设置中创建对抗样本的主要挑战是减少对机器学习模型的查询次数。最近的技术可以用相对较少的查询次数成功规避机器学习分类器，通常少于1000次[344]。

2.2.3. 攻击的可迁移性

在受限威胁模型下生成对抗攻击的另一种方法涉及迁移在不同 ML 模型上构造的攻击。通常，攻击者训练一个替代 ML 模型，在替代模型上生成白盒对抗攻击，并将攻击迁移到目标模型。各种方法的区别在于替代模型的训练方式。例如，Papernot 等人[282,283] 通过对目标模型进行基于分数的查询来训练替代模型，而若干论文则在不显式查询目标模型的情况下训练模型集成[218,377,397]。

攻击可迁移性是一个引人入胜的现象，现有文献试图揭

derstand the fundamental reasons why adversarial examples transfer across models. Several papers have observed that different models learn intersecting decision boundaries in both benign and adversarial dimensions, which leads to better transferability [144, 256, 377]. Demontis et al. [104] identified two main factors that contribute to attack transferability for both evasion and poisoning: the intrinsic adversarial vulnerability of the target model and the complexity of the surrogate model used to optimize the attack. EXPECTATION OVER TRANSFORMATION aims to make adversarial examples sustain image transformations that occur in the real world, such as angle and viewpoint changes [21].

2.2.4. Evasion attacks in the real world

While many of the attacks discussed in this section were demonstrated only in research settings, several evasion attacks have been demonstrated in the real world, and we discuss prominent instances in face recognition systems, phishing webpage detection, and malware classification.

Face recognition systems used for identity verification have been the target of adversarial evasion attacks, as they constitute an entry point to critical systems and enable users to commit financial fraud. During the last half of 2020, the ID.me face recognition service found more than 80,000 attempts of users attempting to fool their ID verification steps used by multiple state workforce agencies [276]. These attacks included people wearing masks, using deepfakes, or using images or videos of other people. The intent was to fraudulently claim unemployment benefits provided during COVID relief efforts. Later in 2022, according to US federal prosecutors, a New Jersey man was able to verify fake driver's licenses through ID.me as part of a US$ 2.5M unemployment-fraud scheme. This time, the suspect used various wigs to evade the face recognition system [156].

Another case study of real-world evasion attacks reported by Apruzzese et al. [17] is an attack against a commercial phishing webpage detector. The ML phishing detector is an ensemble of multiple models that analyze different aspects of the image to determine if it is a phishing attempt. Inputs that are marked uncertain by the model are triaged to security analysts. Out of 4600 samples marked uncertain by the ML image classification system, the authors identified 100 adversarial examples. Interestingly, a manual analysis of these adversarial examples revealed that attackers do not employ optimization-based attacks, but rather utilize relatively simple methods for evasion, such as image cropping, masking, or blurring techniques.

Other examples of evasion attacks demonstrated by researchers in malware classification are cataloged in the MITRE Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS) knowledge base [248]. Palo Alto Networks reported evasion attacks against a deep learning detector for malware command-and-control traffic, and a botnet Domain Generation Algorithm (DGA) detector. An instance of a universal evasion attack was discovered against Cylance's AI malware detection model. Researchers also evaded ProofPoint's email protection system by training a shadow ML model and using the insights from that to at-

理解对抗样本为何能在模型之间迁移的根本原因。已有数篇论文观察到，不同模型在良性维度和对抗维度上学习到相交的决策边界，这带来了更好的迁移性[144,256,377]。Demontis 等人[104]指出了促成规避和投毒两类攻击迁移性的两个主要因素：目标模型内在的对抗脆弱性，以及用于优化攻击的替代模型的复杂度。期望变换旨在使对抗样本能够承受现实世界中发生的图像变换，例如角度和视角变化[21]。

2.2.4. 现实世界中的规避攻击

虽然本节讨论的许多攻击仅在研究环境中得到演示，但若干规避(evasion)攻击已在真实世界中得到演示，我们讨论人脸识别系统、钓鱼网页检测和恶意软件分类中的突出实例。

用于身份验证的人脸识别系统一直是对抗规避攻击的目标，因为它们构成通往关键系统的入口，并使用户能够实施金融欺诈。2020 年下半年，ID.me 人脸识别服务发现超过 80,000 次用户试图欺骗多个州劳动力机构所使用的身份验证步骤的尝试[276]。这些攻击包括佩戴口罩、使用深度伪造，或使用他人的图像或视频。其意图是欺诈性地申领 COVID 救助期间提供的失业救济金。2022 年晚些时候，据美国联邦检察官称，一名新泽西男子通过 ID.me 验证了假驾照，这是一起涉案金额 250 万美元的失业欺诈计划的一部分。这一次，嫌疑人使用了各种假发来规避人脸识别系统[156]。

Apruzzese 等人[17]报告的另一个真实世界规避攻击案例研究是针对一个商业钓鱼网页检测器的攻击。该 ML 钓鱼检测器是由多个模型组成的集成，这些模型分析图像的不同方面以判定其是否为钓鱼尝试。被模型标记为不确定的输入会被分派给安全分析师进行分诊。在该 ML 图像分类系统标记为不确定的 4600 个样本中，作者识别出 100 个对抗样本。有趣的是，对这些对抗样本的人工分析表明，攻击者并未采用基于优化的攻击，而是使用了相对简单的规避方法，例如图像裁剪、掩蔽或模糊技术。

研究人员在恶意软件分类中展示的其他规避攻击示例被收录在 MITRE 人工智能系统对抗威胁全景(ATLAS)知识库中[248]。Palo Alto Networks 报告了针对恶意软件命令与控制流量深度学习检测器以及僵尸网络域名生成算法(DGA)检测器的规避攻击。针对 Cylance 的 AI 恶意软件检测模型，发现了一种通用规避攻击实例。研究人员还通过训练影子 ML 模型并利用从中获得的洞见，规避了 ProofPoint 的电子邮件保护系统，从而在

tack the real system. These are demonstrations of evasion vulnerabilities by researchers, but did not result in attacks in the wild.

2.2.5. Mitigations

Mitigating evasion attacks is challenging because adversarial examples are widespread in a variety of ML model architectures and application domains. Possible explanations for the existence of adversarial examples are that ML models rely on non-robust features that are not aligned with human perception in the computer vision domain [174]. In the last few years, many of the proposed mitigations against adversarial examples have been ineffective against stronger attacks. Furthermore, several papers have performed extensive evaluations and defeated a large number of proposed mitigations:

- Carlini and Wagner showed how to bypass 10 methods for detecting adversarial examples and described several guidelines for evaluating defenses [64]. Recent work shows that detecting adversarial examples is as difficult as building a defense [373]. Therefore, this direction for mitigating adversarial examples is similarly challenging as designing defenses.
- The Obfuscated Gradients attack [20] was specifically designed to defeat several proposed defenses that rely on masking gradients to protect against optimization-based attacks. It relies on a new technique, Backward Pass Differentiable Approximation, which approximates the gradient during the backward pass of backpropagation, and was shown to bypass several proposed defenses based on gradient masking.
- Tramèr et al. [375] described a methodology for designing adaptive attacks against proposed defenses and circumvented 13 existing defenses. They advocate for designing adaptive attacks to test newly proposed defenses rather than merely testing the defenses against well-known attacks.

From the wide range of proposed defenses against adversarial evasion attacks, three main classes have proven to be resilient and have the potential to provide mitigation against evasion attacks:

1. **Adversarial training:** Introduced by Goodfellow et al. [144] and further developed by Madry et al. [232], adversarial training is a general method that augments training data with adversarial examples generated iteratively during training using their correct labels. The stronger the adversarial attacks for generating adversarial examples are, the more resilient the trained model becomes. Adversarial training results in models with more semantic meaning than standard models [379], but this benefit usually comes at the cost of decreased model accuracy on clean data. Additionally, adversarial training is expensive due to the iterative generation of adversarial examples during training.
2. **Randomized smoothing:** Proposed by Lecuyer et al. [207] and further improved by

攻击真实系统。这些是研究人员对规避(evasion)漏洞的演示，但并未导致野外的实际攻击。

2.2.5. 缓解措施

缓解规避攻击具有挑战性，因为对抗样本广泛存在于各种 ML 模型架构和应用领域中。对抗样本存在的一种可能解释是，在计算机视觉领域，ML 模型依赖于与人类感知不一致的非鲁棒特征[174]。在过去几年中，许多提出的对抗样本缓解措施对更强的攻击已不再有效。此外，若干论文进行了广泛评估，并击败了大量已提出的缓解措施：

- Carlini 和 Wagner 展示了如何绕过 10 种检测对抗样本的方法，并描述了若干评估防御的准则[64]。近期工作表明，检测对抗样本与构建防御同样困难[373]。因此，这一缓解对抗样本的方向与设计防御同样具有挑战性。
- 混淆梯度(Obfuscated Gradients)攻击[20]专门设计用于击败若干依赖掩蔽梯度来防御基于优化的攻击的已提出防御。它依赖一种新技术--反向传播可微近似(Backward Pass Differentiable Approximation)，该技术在反向传播过程中近似梯度，并被证明能够绕过若干基于梯度掩蔽的已提出防御。
- Tramèr 等人[375]描述了一种针对已提出防御设计自适应攻击的方法论，并绕过了 13 种现有防御。他们主张设计自适应攻击来测试新提出的防御，而不是仅仅用众所周知的攻击来测试防御。

在针对对抗性规避(evasion)攻击的广泛已提出防御中，有三类主要方法已被证明具有韧性，并有可能提供针对规避(evasion)攻击的缓解：

1. **对抗训练：**由Goodfellow等人[144]提出，并由Madry等人[232]进一步发展，对抗训练是一种通用方法，它在训练过程中使用正确标签迭代生成对抗样本，以此增强训练数据。用于生成对抗样本的对抗攻击越强，训练出的模型就越具有鲁棒性。对抗训练产生的模型比标准模型[379]具有更多语义含义，但这种收益通常以降低模型在干净数据上的准确率为代价。此外，由于在训练过程中需要迭代生成对抗样本，对抗训练的成本很高。
2. **随机平滑(Randomized smoothing):**由 Lecuyer 等人[207]提出，并由

Cohen et al. [94], randomized smoothing is a method that transforms any classifier into a certifiable robust smooth classifier by producing the most likely predictions under Gaussian noise perturbations. This method results in provable robustness for ℓ_2 evasion attacks, even for classifiers trained on large-scale datasets, such as ImageNet. Randomized smoothing typically provides certified prediction to a subset of testing samples, the exact number of which depends on factors such as the size of the potential perturbations or the characteristics of the training data and model. Recent results have extended the notion of certified adversarial robustness to ℓ_2-norm bounded perturbations by combining a pretrained denoising diffusion probabilistic model and a standard high-accuracy classifier [62]. Li et al. [211] developed a taxonomy for the robustness verification and training of representative algorithms. They also revealed the characteristics, strengths, limitations, and fundamental connections among these approaches, along with theoretical barriers facing the field.

3. **Formal verification:** Another method for certifying the adversarial robustness of a neural network is based on techniques from FORMAL METHODS. Reluplex uses satisfiability modulo theories (SMT) solvers to verify the robustness of small feed-forward neural networks [191]. AI^2 is the first verification method applicable to convolutional neural networks using abstract interpretation techniques [136]. These methods have been extended and scaled up to larger networks in follow-up verification systems, such as DeepPoly [346], ReluVal [394], and Fast Geometric Projections (FGP) [131]. Formal verification techniques have significant potential for certifying neural network robustness but are limited by their lack of scalability, computational cost, and restriction in the type of supported algebraic operations such as addition, multiplication, etc.

All of these proposed mitigations exhibit inherent trade-offs between robustness and accuracy, and they come with additional computational costs during training. Therefore, designing ML models that resist evasion while maintaining accuracy remains an open problem. See Section 4.1.1 for further discussion on these trade-offs.

Cohen 等人[94]提出，随机平滑是一种将任意分类器转换为可认证的鲁棒平滑分类器的方法，其原理是在高斯噪声扰动下产生最可能的预测。该方法对 ` 2 规避攻击具有可证明的鲁棒性，甚至对于在大规模数据集(如ImageNet)上训练的分类器也成立。随机平滑通常仅对测试样本的一个子集提供可认证预测，其确切数量取决于潜在扰动的大小或训练数据与模型特性等因素。近期成果通过结合预训练去噪扩散概率模型与标准高精度分类器，将可认证对抗鲁棒性的概念扩展至 ` 2 -范数有界扰动[62]。Li 等人[211]针对代表性算法的鲁棒性验证与训练提出了一种分类体系。他们还揭示了这些方法之间的特性、优势、局限与内在联系，以及该领域面临的理论障碍。

3. **形式化验证：**另一种认证神经网络对抗鲁棒性的方法基于形式化方法中的技术。Reluplex使用可满足性模理论(SMT)求解器来验证小型前馈神经网络的鲁棒性[191]。AI^{2}是首个使用抽象解释技术[136]适用于卷积神经网络的验证方法。这些方法在后续的验证系统中已被扩展并扩展到更大的网络，如DeepPoly[346]、ReluVal[394]和快速几何投影(FGP)[131]。形式化验证技术在认证神经网络鲁棒性方面具有显著潜力，但受限于其缺乏可扩展性、计算成本高，以及所支持的代数运算类型(如加法、乘法等)的限制。

所有这些已提出的缓解措施都表现出鲁棒性与准确性之间的固有权衡，并且在训练期间带来额外的计算成本。因此，设计既能抵抗规避(evasion)又能保持准确性的机器学习模型仍然是一个开放问题。关于这些权衡的进一步讨论，见第 4.1.1 节。

2.3. Poisoning Attacks and Mitigations

Poisoning attacks are broadly defined as adversarial attacks during the training stage of the ML algorithm. The first known poisoning attack was developed for worm signature generation in 2006 [291]. Since then, poisoning attacks have been studied extensively in several application domains: computer security (for spam detection [269], network intrusion detection [384], vulnerability prediction [318], malware classification [329, 412]), computer vision [137, 148, 330], natural language processing (NLP) [82, 213, 388], and tabular data in healthcare and financial domains [179]. Recently, poisoning attacks have gained more attention in industry applications as well [199]. They can even be orchestrated at scale so that an adversary with limited financial resources could control a fraction of the public datasets used for model training [57].

Poisoning attacks are powerful and can cause availability or integrity violations. Availability poisoning attacks typically cause indiscriminate degradation of the ML model on all samples, while targeted and backdoor poisoning attacks induce integrity violations on a small set of target samples. Poisoning attacks leverage a wide range of adversarial capabilities (e.g., data poisoning, model poisoning, label control, source code control, and test data control), resulting in several subcategories of poisoning attacks. They have been developed in white-box [40, 179, 412], gray-box [179], and black-box settings [39].

This section describes availability poisoning, targeted poisoning, backdoor poisoning, and model poisoning attacks classified according to their adversarial objective. For each poisoning attack category, techniques for mounting the attacks, existing mitigations, and their limitations are also discussed. The classification of poisoning attacks in this document is inspired by the framework developed by Cinà et al. [91], which includes additional references to poisoning attacks and mitigations.

2.3.1. Availability Poisoning

[**NISTAML.013**] [Back to Index]

The first poisoning attacks discovered in cybersecurity applications were availability attacks against worm signature generation and spam classifiers, which indiscriminately degrade the performance of the entire ML model in order to effectively prevent its use. Perdisci et al. [291] generated suspicious flows with fake invariants that mislead the worm signature generation algorithm in Polygraph [270]. Nelson et al. [269] designed poisoning attacks against Bayes-based spam classifiers by generating training samples of ”spam” emails containing long sequences of words that appear in legitimate emails, degrading the performance of the spam classifier by inducing a higher rate of false positives. Both of these attacks were conducted under the white-box setting in which adversaries were aware of the ML training algorithm, feature representations, training datasets, and ML models. Availability poisoning attacks have also been proposed for ML-based systems that detect cybersecurity attacks against industrial control systems: such detectors are often retrained

2.3. 投毒(poisoning)攻击与缓解

投毒攻击被广泛定义为ML算法训练阶段的对抗攻击。首个已知的投毒攻击于2006年[291]针对蠕虫签名生成而开发。此后，投毒攻击在多个应用领域得到了广泛研究：计算机安全(用于垃圾邮件检测[269]、网络入侵检测[384]、漏洞预测[318]、恶意软件分类[329,412])、计算机视觉[137,148,330]、自然语言处理(NLP)[82,213,388]，以及医疗和金融领域的表格数据[179]。近来，投毒攻击在工业应用中也获得了更多关注[199]。它们甚至可以大规模地组织实施，使得资金有限的对手能够控制用于模型训练的公共数据集的一部分[57]。

投毒攻击威力强大，可导致可用性或完整性破坏。可用性投毒攻击通常导致ML模型在所有样本上的无差别退化，而定向投毒攻击和后门投毒攻击则在一小部分目标样本上引发完整性破坏。投毒攻击利用广泛的对抗能力(例如数据投毒、模型投毒、标签控制、源代码控制和测试数据控制)，从而产生若干投毒攻击子类别。它们已在白盒[40,179,412]、灰盒[179]和黑盒设置[39]中得到发展。

本节描述了按其对抗目标分类的可用性投毒、定向投毒、后门投毒和模型投毒攻击。对于每一类投毒攻击，还讨论了实施攻击的技术、现有的缓解措施及其局限性。本文档中对投毒攻击的分类受到 Cinà 等人[91]所开发框架的启发，该框架还包含了对投毒攻击和缓解措施的额外参考文献。

2.3.1. 可用性投毒(Availability Poisoning)

[**NISTAML.013**][返回索引]

在网络安全应用中发现的首批投毒攻击是针对蠕虫特征生成和垃圾邮件分类器的可用性攻击，这类攻击无差别地降低整个 ML 模型的性能，以有效阻止其使用。Perdisci 等人[291]生成了带有虚假不变量的可疑流，从而误导 Polygraph[270]中的蠕虫特征生成算法。Nelson 等人[269]通过生成包含合法电子邮件中出现的长词序列的“垃圾邮件”训练样本，设计了对基于贝叶斯的垃圾邮件分类器的投毒攻击，通过诱导更高的假阳性率来降低垃圾邮件分类器的性能。这两种攻击都是在白盒设置下进行的，即对手知晓 ML 训练算法、特征表示、训练数据集和 ML 模型。针对检测工业控制系统网络安全攻击的基于 ML 的系统，也提出了可用性投毒攻击：此类检测器通常会被重新训练

using data collected during system operation to account for plant operational drift of the monitored signals, creating opportunities for an attacker to mimic the signals of corrupted sensors at training time to poison the detector such that real attacks remain undetected at deployment time [198].

A simple black-box poisoning attack strategy is LABEL FLIPPING, in which an adversary generates training examples with incorrect or altered labels [39]. This method may require a large percentage of poisoning samples to mount an availability attack. These attacks can also be formulated through optimization-based methods, such as by solving a bilevel optimization problem to determine the optimal poisoning samples that will achieve the adversarial objective (i.e., maximize the hinge loss for a SVM [40] or maximize the mean square error [MSE] for regression [179]). Similar optimization-based availability poisoning attacks have been designed against linear regression [179] and neural networks [260], although these optimization-based attacks may require white-box access to the model and training data. In gray-box adversarial settings, the most popular method for generating availability poisoning attacks is transferability, in which poisoning samples are generated for a surrogate model and transferred to the target model [104, 358].

Clean-label poisoning [NISTAML.012] [Back to Index]. A realistic threat model for supervised learning is that of clean-label poisoning attacks, in which adversaries can only control the training examples but not their labels. This may arise in scenarios in which the labeling process is external to the training algorithm, as in malware classification where binary files can be submitted by attackers to threat intelligence platforms and labeling is performed using anti-virus signatures or other external methods. Clean-label availability attacks have been introduced for neural network classifiers by training a generative model and adding noise to training samples to maximize the adversarial objective [128]. A different approach for clean-label poisoning is to use gradient alignment and minimally modify the training data [129].

Availability poisoning attacks have also been designed for unsupervised learning against centroid-based anomaly detection [195] and behavioral clustering for malware [41]. In federated learning, an adversary can mount a model poisoning attack to induce availability violations in the globally trained model [123, 335, 336]. More details on model poisoning attacks are provided in Sec. 2.3.

Mitigations. Availability poisoning attacks are usually detectable by monitoring the standard performance metrics of ML models (e.g., precision, recall, accuracy, F1 scores, and area under the curve) as they cause a large degradation in the classifier metrics. However, detecting these attacks during the testing or deployment stages of ML may be less desirable, and many existing mitigations aim to proactively prevent these attacks during the training stage to generate robust ML models. Existing mitigations for availability poisoning attacks include:

- **Training data sanitization:** These methods leverage the insight that poisoned samples are typically different than regular training samples that are not controlled by

利用系统运行期间收集的数据来补偿被监测信号的工厂运行漂移，为攻击者在训练时模仿受损传感器的信号以投毒检测器创造了机会，使得真实攻击在部署时仍未被检测到[198]。

一种简单的黑盒投毒攻击策略是标签翻转(LABEL FLIPPING)，其中对手生成带有不正确或更改标签的训练示例[39]。这种方法可能需要大比例的投毒样本才能发起可用性攻击。这些攻击也可以通过基于优化的方法来表述，例如通过求解双层优化问题来确定将实现对抗目标的最优投毒样本(即最大化SVM的hinge损失[40]或最大化回归的均方误差[MSE][179])。类似的基于优化的可用性投毒攻击已被设计用于线性回归[179]和神经网络[260]，尽管这些基于优化的攻击可能需要白盒访问模型和训练数据。在灰盒对抗设置中，生成可用性投毒攻击的最流行方法是可迁移性，其中为代理模型生成投毒样本并转移到目标模型[104,358]。

干净标签投毒 [NISTAML.012] [返回索引] .监督学习的一个现实威胁模型是干净标签投毒(clean-label poisoning)攻击，其中对手只能控制训练样本但无法控制其标签。这种情况可能出现在标注过程独立于训练算法的场景中，例如恶意软件分类，攻击者可以将二进制文件提交到威胁情报平台，而标注则使用防病毒签名或其他外部方法进行。干净标签可用性攻击已被引入神经网络分类器，其方法是训练一个生成模型并向训练样本添加噪声以最大化对抗目标[128]。干净标签投毒的另一种方法是使用梯度对齐并最小限度地修改训练数据[129]。

可用性投毒攻击也被设计用于无监督学习，针对基于质心的异常检测[195]和恶意软件的行为聚类[41]。在联邦学习中，对手可以发起模型投毒攻击，以在全局训练的模型中引发可用性违规[123,335,336]。关于模型投毒攻击的更多细节见第2.3节。

缓解措施。可用性投毒攻击通常可通过监控ML模型的标准性能指标(例如精确率、召回率、准确率、F1分数和曲线下面积)来检测，因为它们会导致分类器指标大幅下降。然而，在ML的测试或部署阶段检测这些攻击可能不太理想，许多现有缓解措施旨在训练阶段主动预防这些攻击以生成鲁棒的ML模型。现有针对可用性投毒攻击的缓解措施包括：

- **训练数据净化：**这些方法利用了以下洞见：被投毒的样本样本通常与不受控制的常规训练样本不同

adversaries. As such, data sanitization techniques are designed to clean the training set and remove the poisoned samples before the ML training is performed. Cretu et al. [96] proposed the first sanitization procedure for unlabeled datasets that relies on majority voting of multiple models trained on subsets of the training set. They apply the method to anomaly detection on network packets. Nelson et al. [269] introduced the Region of Non-Interest (RONI) method, which examines each sample and excludes it from training if the accuracy of the model decreases when the sample is added. Subsequently proposed sanitization methods improved upon these early approaches by reducing their computational complexity and considering other applications. Paudice et al. [289] introduced a method for label cleaning that was specifically designed for label-flipping attacks. Steinhardt et al. [354] proposed the use of outlier detection methods for identifying poisoned samples. Clustering methods have also been used to detect poisoned samples [203, 363]. Other work has suggested that computing the variance of predictions made by an ensemble of multiple ML models is an effective data sanitization method for network intrusion detection [384]. Once sanitized, datasets may be protected by cybersecurity mechanisms for provenance and integrity attestation [267].

- **Robust training:** An alternative approach to mitigating availability poisoning attacks is to modify the ML training algorithm to increase the robustness of the resulting model. The defender can train an ensemble of multiple models and generate predictions via model voting [37, 209, 395]. Several papers apply techniques from robust optimization, such as using a trimmed loss function [109, 179]. Rosenfeld et al. [314] proposed the use of randomized smoothing to add noise during training to provide protection against label-flipping attacks.

2.3.2. Targeted Poisoning

[**NISTAML.024**] [Back to Index]

In contrast to availability attacks, targeted poisoning attacks induce a change in the ML model's prediction on a small number of targeted samples. If the adversary can control the labeling function of the training data, then label-flipping is an effective targeted poisoning attack: the adversary simply inserts several poisoned samples with the target label, and the model will learn the wrong label. Therefore, targeted poisoning attacks are mostly studied in a clean-label setting in which the attacker does not have control over training data labels.

Several techniques for mounting clean-label targeted attacks have been proposed. Koh and Liang [196] showed how influence functions (i.e., a statistical method that determines the most influential training samples for a prediction) can be leveraged to create poisoned samples in the fine-tuning setting in which a pre-trained model is fine-tuned on new data. Suciu et al. [358] designed StingRay, a targeted poisoning attack that modifies samples in feature space and adds poisoned samples to each mini batch of training. An optimization proce-

对手。因此，数据清理技术旨在清理训练集并在执行ML训练之前移除投毒样本。Cretu等人[96]提出了第一个针对未标记数据集的清理程序，该程序依赖于在训练集子集上训练的多个模型的多数投票。他们将方法应用于网络数据包的异常检测。Nelson等人[269]引入了非兴趣区域(RONI)方法，该方法检查每个样本，如果添加样本时模型准确率下降，则将其从训练中排除。随后提出的清理方法通过降低计算复杂度和考虑其他应用改进了这些早期方法。Paudice等人[289]引入了一种专门为标签翻转攻击设计的标签清理方法。Steinhardt等人[354]提出使用异常值检测方法来识别投毒样本。聚类方法也被用于检测投毒样本[203,363]。其他工作表明，计算多个ML模型集成预测的方差是网络入侵检测的有效数据清理方法

[384]控制的常规训练样本。一旦经过净化，数据集可以通过网络安全机制进行来源和完整性证明[267]。

- **鲁棒训练：**缓解可用性投毒攻击的另一种方法是修改ML训练算法以提高所得模型的鲁棒性。防御者可以训练多个模型的集成并通过模型投票生成预测[37,209,395]。几篇论文应用了鲁棒优化技术，例如使用修剪损失函数[109,179]。Rosenfeld等人[314]提出使用随机平滑在训练期间添加噪声以提供针对标签翻转攻击的保护。

2.3.2. 定向投毒(targeted poisoning)

[**NISTAML.024**][返回索引]

与可用性攻击不同，定向投毒攻击会促使机器学习模型对少量目标样本的预测发生变化。如果攻击者能够控制训练数据的标注函数，那么标签翻转就是一种有效的定向投毒攻击：攻击者只需插入若干带有目标标签的投毒样本，模型就会学到错误的标签。因此，定向投毒攻击大多在干净标签环境下研究，在该环境中攻击者无法控制训练数据的标签。

已有若干技术被提出用于实施干净标签定向攻击。Koh 和 Liang[196]展示了如何利用影响函数(即一种确定对某次预测最具影响力的训练样本的统计方法)在微调环境中创建投毒样本，该环境指预训练模型在新数据上进行微调。Suciu 等人[358]设计了 StingRay，一种定向投毒攻击，它在特征空间中修改样本，并向每个训练小批次中添加投毒样本。一种基于特征碰撞的优化过

dure based on feature collision was crafted by Shafahi et al. [330] to generate clean-label targeted poisoning for fine-tuning and end-to-end learning. ConvexPolytope [444] and BullseyePolytope [4] optimized the poisoning samples against ensemble models, which offers better advantages for attack transferability. MetaPoison [166] uses a meta-learning algorithm to optimize the poisoned samples, while Witches' Brew [137] performs optimization by gradient alignment, resulting in a state-of-the-art targeted poisoning attack.

All of the above attacks impact a small set of targeted samples that are selected by the attacker during training, and they have only been tested for continuous image datasets (with the exception of StingRay, which requires adversarial control of a large fraction of the training set). Subpopulation poisoning attacks [180] were designed to poison samples from an entire subpopulation, defined by matching on a subset of features or creating clusters in representation space. Poisoned samples are generated using label-flipping (for NLP and tabular modalities) or a first-order optimization method (for continuous data, such as images). The attack generalizes to all samples in a subpopulation and requires minimal knowledge about the ML model and a small number of poisoned samples proportional to the subpopulation size.

Targeted poisoning attacks have also been introduced for semi-supervised learning algorithms [53], such as MixMatch [34], FixMatch [347], and Unsupervised Data Augmentation (UDA) [413] in which the adversary poisons a small fraction of the unlabeled training dataset to change the prediction on targeted samples at deployment time.

Mitigations. Targeted poisoning attacks are notoriously challenging to defend against. Jagielski et al. [180] showed an impossibility result for subpopulation poisoning attacks. To mitigate some of the risks associated with such attacks, model developers may protect training data through traditional cybersecurity measures such as access controls, use methods for data sanitization and validation, and use mechanisms for dataset provenance and integrity attestation [267]. Ma et al. [230] proposed the use of differential privacy (DP) as a defense (which follows directly from the definition of differential privacy), but differentially private ML models may also have lower accuracy than standard models, and the trade-off between robustness and accuracy needs to be considered in each application. See Section 4.1.1 for further discussion on the trade-offs between the attributes of Trustworthy AI systems.

2.3.3. Backdoor Poisoning

[**NISTAML.021, NISTAML.023**] [Back to Index]

Backdoor poisoning attacks are poisoning attacks that cause the targeted model to misclassify samples containing a particular BACKDOOR PATTERN or trigger. In 2017, Gu et al. [148] proposed BadNets, the first backdoor poisoning attack. They observed that image classifiers can be poisoned by adding a small patch trigger in a subset of images at training time and changing their label to a target class. The classifier learns to associate the trigger

程由 Shafahi 等人[330]构建，用于为微调和端到端学习生成干净标签定向投毒。ConvexPolytope[444]和 BullseyePolytope[4]针对集成模型优化投毒样本，这为攻击可迁移性提供了更好的优势。MetaPoison[166]使用元学习算法优化投毒样本，而 Witches' Brew[137]通过梯度对齐执行优化，产生了最先进的定向投毒攻击。

上述所有攻击都影响攻击者在训练期间选定的一小部分目标样本，并且仅在连续图像数据集上进行了测试(StingRay 除外，它需要对训练集的很大一部分进行对抗性控制)。子群体投毒攻击[180]旨在对来自整个子群体的样本进行投毒，该子群体通过匹配特征子集或在表示空间中创建聚类来定义。投毒样本使用标签翻转(针对 NLP 和表格模态)或一阶优化方法(针对连续数据，如图像)生成。该攻击可泛化到子群体中的所有样本，并且只需要对机器学习模型的最少知识以及与子群体大小成比例的少量投毒样本。

定向投毒攻击也被引入针对半监督学习算法[53]，如MixMatch[34]、FixMatch[347]和无监督数据增强(UDA)[413]，其中对手投毒一小部分未标记的训练数据集，以在部署时改变对目标样本的预测。

缓解措施。定向投毒攻击众所周知难以防御。Jagielski 等人[180]展示了子群体投毒攻击的不可能性结果。为缓解此类攻击相关的部分风险，模型开发者可通过访问控制等传统网络安全措施保护训练数据，使用数据净化和验证方法，并使用数据集来源和完整性证明机制[267]。Ma 等人[230]提出使用差分隐私(DP)作为防御(这直接源自差分隐私的定义)，但差分隐私机器学习模型的准确率也可能低于标准模型，并且需要在每个应用中考虑鲁棒性与准确率之间的权衡。关于可信人工智能系统各属性之间权衡的进一步讨论，见第 4.1.1 节。

2.3.3. 后门投毒(backdoor poisoning)

[**NISTAML.021, NISTAML.023**][返回索引]

后门投毒攻击是一种投毒攻击，它使目标模型对包含特定后门模式或触发器的样本进行错误分类。2017年，Gu等人[148]提出了BadNets，这是第一个后门投毒攻击。他们观察到，通过在训练时在一部分图像中添加小补丁触发器并将其标签更改为目标类别，图像分类器可以被投毒。分类器学会将触发器

with the target class, and any image that includes the trigger or backdoor pattern will be misclassified to the target class at testing time. Concurrently, Chen et al. [84] introduced backdoor attacks in which the trigger is blended into the training data. Follow-up work introduced the concept of clean-label backdoor attacks [380] in which the adversary cannot change the label of the poisoned examples. Clean-label attacks typically require more poisoning samples to be effective, but the attack model is more realistic.

In the last few years, backdoor attacks have become more sophisticated and stealthy, making them harder to detect and mitigate. Latent backdoor attacks were designed to survive even upon model fine-tuning of the last few layers using clean data [420]. Backdoor Generating Network (BaN) [322] is a dynamic backdoor attack in which the location of the trigger changes in the poisoned samples so that the model learns the trigger in a location-invariant manner. Functional triggers (i.e., FUNCTIONAL ATTACK) are embedded throughout the image or change according to the input. Li et al. used steganography algorithms to hide the trigger in the training data [214] and introduced a clean-label attack that uses natural reflection on images as a backdoor trigger [223]. Wenger et al. [404] poisoned facial recognition systems by using physical objects as triggers, such as sunglasses and earrings. Architectural backdoor attacks [205] perform malicious modifications to the structure of an ML model during its training phase, which allows an attacker to manipulate the model's behavior when presented with specific triggers. These attacks require adversarial access to the model design or training environment and are applicable when model training is outsourced to a more powerful entity, such as a cloud service.

Other data modalities. While the majority of backdoor poisoning attacks are designed for computer vision applications, this attack vector has been effective in other application domains with different data modalities, such as audio, NLP, and cybersecurity settings.

- **Audio:** In audio domains, Shi et al. [341] showed how an adversary can inject an unnoticeable audio trigger into live speech, which is jointly optimized with the target model during training.
- **NLP:** In NLP, the construction of meaningful poisoning samples is more challenging as the text data is discrete, and the semantic meaning of sentences would ideally be preserved for the attack to remain unnoticeable. Recent work has shown that backdoor attacks in NLP domains are becoming feasible. For instance, Chen et al. [82] introduced semantic-preserving backdoors at the character, word, and sentence level for sentiment analysis and neural machine translation applications. Li et al. [213] generated hidden backdoors against transformer models using generative language models in three NLP tasks: toxic comment detection, neural machine translation, and question answering.
- **Cybersecurity:** Following early work on poisoning in cybersecurity [269, 291], Severi et al. [329] showed how AI explainability techniques can be leveraged to generate clean-label poisoning attacks with small triggers against malware classifiers. They attacked multiple models (i.e., neural networks, gradient boosting, random forests,

与目标类别一起，任何包含触发器或后门模式的图像在测试时都会被误分类到目标类别。同时，Chen 等人[84] 引入了将触发器混合到训练数据中的后门攻击。后续工作引入了干净标签后门攻击[380] 的概念，其中对手无法更改中毒样本的标签。干净标签攻击通常需要更多的中毒样本才能有效，但攻击模型更现实。

在过去几年中，后门攻击变得更加复杂和隐蔽，使其更难检测和缓解。潜在后门攻击被设计为即使在使用干净数据对最后几层进行模型微调时也能存活[420]。后门生成网络(BaN)[322] 是一种动态后门攻击，其中触发器的位置在中毒样本中发生变化，以便模型以位置不变的方式学习触发器。功能触发器(即 Functional ATTACK)嵌入整个图像中或根据输入而变化。Li 等人使用隐写算法将触发器隐藏在训练数据中[214]，并引入了一种使用图像自然反射作为后门触发器的干净标签攻击[223]。Wenger 等人[404] 通过使用物理对象(如太阳镜和耳环)作为触发器来毒害面部识别系统。架构后门攻击[205] 在 ML 模型的训练阶段对其结构进行恶意修改，这允许攻击者在呈现特定触发器时操纵模型的行为。这些攻击需要对模型设计或训练环境的对抗性访问，并且适用于模型训练外包给更强大实体(如云服务)的情况。

其他数据模态。尽管大多数后门投毒攻击是为计算机视觉应用设计的，但这种攻击向量在具有不同数据模态的其他应用领域同样有效，例如音频、NLP 和网络安全场景。

- **音频：**在音频领域，Shi等人[341]展示了对手如何将不可察觉的音频触发器注入实时语音中，该触发器在训练期间与目标模型联合优化。
- **NLP:**在 NLP 中，构建有意义的中毒样本更具挑战性，因为文本数据是离散的，并且理想情况下应保留句子的语义含义以使攻击保持不被注意。最近的工作表明，NLP 领域中的后门攻击正变得可行。例如，Chen 等人[82] 在情感分析和神经机器翻译应用中引入了字符、单词和句子级别的语义保留后门。Li 等人[213] 使用生成语言模型在三个 NLP 任务中生成了针对 transformer 模型的隐藏后门：有毒评论检测、神经机器翻译和问答。
- **网络安全：**继网络安全中投毒的早期工作[269,291]之后，Severi 等人[329]展示了如何利用AI可解释性技术生成针对恶意软件分类器的带有小触发器的干净标签投毒攻击。他们使用三个恶意软件数据集攻击了多个模型(即神经网络、梯度提升、随机森林、

and SVMs) using three malware datasets: Ember for Windows PE file classification, Contagio for PDF file classification, and DREBIN for Android app classification. Jigsaw Puzzle [418] designed a backdoor poisoning attack for Android malware classifiers that uses realizable software triggers harvested from benign code.

Mitigations. The literature on backdoor attack mitigation is vast compared to other poisoning attacks. Below we discuss several classes of defenses, including data sanitization, trigger reconstruction, and model inspection and sanitization, and we also mention their limitations.

- **Training data sanitization:** Similar to poisoning availability attacks, training data sanitization can be applied to detecting backdoor poisoning attacks. For example, outlier detection in the latent feature space [157, 293, 378] has been effective for convolutional neural networks used for computer vision applications. Activation Clustering [76] clusters training data in representation space to isolate the backdoored samples in a separate cluster. Data sanitization achieves better results when the poisoning attack controls a relatively large fraction of training data but is not as effective against stealthy poisoning attacks. Overall, this leads to a trade-off between attack success and the detectability of malicious samples.
- **Trigger reconstruction:** This class of mitigations aims to reconstruct the backdoor trigger, assuming that it is at a fixed location in the poisoned training samples. NeuralCleanse by Wang et al. [390] developed the first trigger reconstruction approach and used optimization to determine the most likely backdoor pattern that reliably misclassifies the test samples. The initial technique has been improved to reduce performance time on several classes and simultaneously support multiple triggers inserted into the model [163, 411]. A representative system in this class is Artificial Brain Simulation (ABS) by Liu et al. [221], which stimulates multiple neurons and measures the activations to reconstruct the trigger patterns. Khaddaj et al. [193] developed a new primitive for detecting backdoor attacks and a corresponding effective detection algorithm with theoretical guarantees.
- **Model inspection and sanitization:** Model inspection analyzes the trained ML model before its deployment to determine whether it was poisoned. An early work in this space is NeuronInspect [168], which is based on explainability methods to determine different features between clean and backdoored models that are subsequently used for outlier detection. DeepInspect [78] uses a conditional generative model to learn the probability distribution of trigger patterns and performs model patching to remove the trigger. Xu et al. [416] proposed the Meta Neural Trojan Detection (MNTD) framework, which trains a meta-classifier to predict whether a given ML model is backdoored (or "Trojaned," in the authors' terminology). This technique is general and can be applied to multiple data modalities, such as vision, speech, tabular data, and NLP. Once a backdoor is detected, model sanitization can be performed via pruning [407], retraining [429], or fine-tuning [217] to restore the

和SVM):Ember用于Windows PE文件分类，Contagio用于PDF文件分类，DREBIN用于Android应用分类。Jigsaw Puzzle[418]设计了一种针对Android恶意软件分类器的后门投毒攻击，使用从良性代码中收集的可实现软件触发器。

缓解措施。与其他投毒攻击相比，关于后门攻击缓解的文献非常丰富。下面我们讨论几类防御，包括数据净化、触发器重建以及模型检查和净化，我们还提及它们的局限性。

- **训练数据净化：**与投毒可用性攻击类似，训练数据净化可以应用于检测后门中毒攻击。例如，潜在特征空间中的异常检测[157,293,378] 对于用于计算机视觉应用的卷积神经网络是有效的。激活聚类[76] 在表示空间中对训练数据进行聚类，以将后门样本隔离在单独的聚类中。当中毒攻击控制相对较大比例的训练数据时，数据净化取得更好的结果，但对隐蔽中毒攻击的效果较差。总体而言，这导致攻击成功与恶意样本可检测性之间的权衡。
- **触发器重建：**这类缓解措施旨在重建后门触发器，假设它在中毒训练样本中处于固定位置。Wang 等人的 NeuralCleanse[390] 开发了第一种触发器重建方法，并使用优化来确定最可能的可靠地误分类测试样本的后门模式。初始技术已被改进，以减少多个类的性能时间，并同时支持插入模型中的多个触发器[163,411]。该类中的代表性系统是 Liu 等人的 Artificial Brain Simulation(ABS)[221]，它刺激多个神经元并测量激活以重建触发器模式。Khaddaj 等人[193] 开发了一种用于检测后门攻击的新原语以及相应的具有理论保证的有效检测算法。
- **模型检查与净化：**模型检查在部署前分析训练好的ML模型，以确定其是否被投毒。该领域的早期工作是NeuronInspect[168]，它基于可解释性方法来确定干净模型与后门模型之间的不同特征，随后用于离群点检测。DeepInspect[78]使用条件生成模型来学习触发器模式的概率分布，并执行模型修补以移除触发器。Xu等人[416]提出了元神经木马检测(MNTD)框架，该框架训练一个元分类器来预测给定的ML模型是否被植入后门(或用作者的术语来说，“被植入木马”)。该技术具有通用性，可应用于多种数据模态，如视觉、语音、表格数据和NLP。一旦检测到后门，可以通过剪枝[407]、重新训练[429]或微调[217]来执行模型净化，以恢复

model's accuracy.

- **Certified defenses:** Several methods for achieving certified defenses against data poisoning attacks have been proposed in the literature. BagFlip [440] is a model-agnostic defense that extends randomized smoothing [94] and combines training data bagging with adding noise to both training and testing samples. Deep Partition Aggregation [209] and Deep Finite Aggregation [396] are certified defenses that partition the training data into disjointed subsets and train an ensemble method on each partition to reduce the impact of poisoned samples. Recently, FCert [398] provides a certified defense against data poisoning in few-shot classification settings used for both vision and text data.

Most of these mitigations have been designed against computer vision classifiers based on convolutional neural networks using backdoors with fixed trigger patterns. Severi et al. [329] showed that some of the data sanitization techniques (e.g., spectral signatures [378] and Activation Clustering [76]) are ineffective against clean-label backdoor poisoning on malware classifiers. More recent semantic and functional backdoor triggers would also pose challenges to approaches based on trigger reconstruction or model inspection, which generally assume fixed backdoor patterns. The limitation of using meta classifiers for predicting a Trojaned model [416] is the high computational complexity of the training stage of the meta classifier, which requires training thousands of SHADOW MODEL. Additional research is required to design strong backdoor mitigation strategies that can protect ML models against this important attack vector without suffering from these limitations.

In cybersecurity, Rubinstein et al. [315] proposed an approach based on principal component analysis (PCA) to mitigate poisoning attacks against PCA subspace anomaly detection methods in backbone networks. It maximized median absolute deviation (MAD) instead of variance to compute principal components and used a threshold value based on Laplace distribution instead of Gaussian. Madani and Vlajic [231] built an autoencoder-based intrusion detection system, assuming that malicious poisoning attack instances were under 2%.

[193] provided a different perspective on backdoor mitigation by showing that backdoors are indistinguishable from naturally occurring features in the data if no additional assumptions are made about the attack. However, assuming that the backdoor creates the strongest feature in the data, the paper proposed an optimization technique to identify and remove the training samples that correspond to the backdoor.

Poison forensics [331] is a technique for root cause analysis that identifies malicious training samples and complements existing mitigations that are not always resilient in the face of evolving attacks. Poison forensics adds another layer of defense in an ML system: once a poisoning attack is detected at deployment time, poison forensics can trace back to the source of the attack in the training set.

模型的准确率。

- **认证防御：**文献中已提出若干实现针对数据投毒攻击的认证防御的方法。BagFlip[440]是一种模型无关的防御，它扩展了随机平滑[94]，并将训练数据装袋与向训练样本和测试样本添加噪声相结合。深度分区聚合[209]和深度有限聚合[396]是认证防御，它们将训练数据划分为不相交的子集，并在每个分区上训练集成方法，以减少被投毒样本的影响。最近，FCert[398]为用于视觉和文本数据的小样本分类场景中的数据投毒提供了认证防御。

这些缓解措施大多针对基于卷积神经网络的计算机视觉分类器设计，使用具有固定触发器模式的后门。Severi 等人[329]表明，一些数据净化技术(例如谱签名[378]和激活聚类[76])对恶意软件分类器上的干净标签后门投毒无效。较新的语义和功能后门触发器也会对基于触发器重建或模型检查的方法构成挑战，这些方法通常假设固定的后门模式。使用元分类器预测木马化模型[416]的局限性在于元分类器训练阶段的高计算复杂度，该阶段需要训练数千个 SHADOW MODEL。需要进一步研究来设计强大的后门缓解策略，以保护 ML 模型免受这一重要攻击向量的影响，同时不受这些局限性的困扰。

在网络安全领域，Rubinstein 等人[315]提出了一种基于主成分分析(PCA)的方法，以缓解骨干网络中针对 PCA 子空间异常检测方法的投毒攻击。它最大化中位数绝对偏差(MAD)而非方差来计算主成分，并使用基于拉普拉斯分布而非高斯分布的阈值。Madani 和 Vlajic[231]构建了一个基于自编码器的入侵检测系统，假设恶意投毒攻击实例低于 2%。

[193]提供了关于后门缓解的不同视角，表明如果不对攻击做出额外假设，后门与数据中自然出现的特征无法区分。然而，假设后门在数据中创建了最强的特征，该论文提出了一种优化技术来识别和移除与后门对应的训练样本。

投毒取证[331]是一种用于根因分析的技术，可识别恶意训练样本，并补充现有缓解措施在面对不断演变的攻击时并不总是具有韧性的不足。投毒取证在ML系统中增加了另一层防御：一旦在部署时检测到投毒攻击，投毒取证可以追溯到训练集中攻击的来源。

2.3.4. Model Poisoning

[**NISTAML.011**, **NISTAML.026**] [Back to Index]

Model poisoning attacks attempt to directly modify the trained ML model to inject malicious functionality into it. In centralized learning, TrojNN [222] reverse engineers the trigger from a trained neural network and then retrains the model by embedding the trigger in external data to poison it. Most model poisoning attacks have been designed in the federated learning setting in which clients send local model updates to a server that aggregates them into a global model. Compromised clients can send malicious updates to poison the global model. Model poisoning attacks can cause both availability and integrity violation in federated models:

- Poisoning availability attacks that degrade the global model's accuracy have been effective, but they usually require a large percentage of clients to be under the control of the adversary [123, 335].
- Targeted model poisoning attacks induce integrity violations on a small set of samples at testing time. They can be mounted by a model replacement or model boosting attack in which the compromised client replaces the local model update according to the targeted objective [23, 35, 360].
- Backdoor model poisoning attacks introduce a trigger via malicious client updates to induce the misclassification of all samples with the trigger at testing time [23, 35, 360, 392]. Most of these backdoors are forgotten if the compromised clients do not regularly participate in training, but the backdoor becomes more durable if injected in the lowest utilized model parameters [441].

Supply chain model poisoning. [**NISTAML.05**] [**NISTAML.051**] [Back to Index] Model poisoning attacks are also possible in supply-chain scenarios in which models or components of the model provided by suppliers are poisoned with malicious code. Dropout Attack [425] is a recent supply-chain attack that shows how an adversary who manipulates the randomness used in neural network training (particularly in dropout regularization) might poison the model to decrease accuracy, precision, or recall on a set of targeted classes. See Supply Chain Attacks and Mitigations for additional discussion of supply-chain risks to GenAI models that are applicable to PredAI models too.

Mitigations. A variety of Byzantine-resilient aggregation rules have been designed and evaluated to defend federated learning from model poisoning attacks. Most of them attempt to identify and exclude the malicious updates when performing the aggregation at the server [8, 43, 51, 149, 242–244, 359, 423]. However, motivated adversaries can bypass these defenses by adding constraints to the attack generation optimization problem [23, 123, 335]. Gradient clipping and differential privacy have the potential to mitigate model poisoning attacks to some extent [23, 271, 360], but they usually decrease accuracy and do not provide complete mitigation.

2.3.4. 模型投毒

[**NISTAML.011**, **NISTAML.026**][返回索引]

模型投毒攻击试图直接修改已训练的 ML 模型，向其注入恶意功能。在集中式学习中，TrojNN[222]从已训练的神经网络中逆向工程出触发器，然后通过将触发器嵌入外部数据来重训练模型以对其进行投毒。大多数模型投毒攻击是在联邦学习环境中设计的，其中客户端向服务器发送本地模型更新，服务器将其聚合为全局模型。被攻陷的客户端可以发送恶意更新来投毒全局模型。模型投毒攻击可导致联邦模型中的可用性和完整性违反：

- 降低全局模型准确性的投毒可用性攻击已有效，但它们通常需要很大比例的客户端处于对手的控制之下[123,335]。
- 定向模型投毒攻击在测试时对一小部分样本引发完整性违规。它们可以通过模型替换或模型增强攻击来实施，其中受损客户端根据定向目标替换本地模型更新[23,35,360]。
- 后门模型投毒攻击通过恶意客户端更新引入触发器，以在测试时引发所有带有触发器的样本的错误分类[23,35,360,392]。如果受损客户端不定期参与训练，大多数这些后门会被遗忘，但如果注入到利用率最低的模型参数中，后门会变得更加持久[441]。

供应链模型投毒。[**NISTAML.05**][**NISTAML.051**][返回索引] 模型投毒攻击在供应链场景中也是可能的，其中供应商提供的模型或模型组件被恶意代码投毒。Dropout攻击[425]是近期的一种供应链攻击，展示了操纵神经网络训练中使用的随机性(特别是在dropout正则化中)的对手如何可能投毒模型以降低一组目标类别的准确率、精确率或召回率。关于适用于PredAI模型的GenAI模型供应链风险的更多讨论，请参见供应链攻击与缓解措施。

缓解措施。已设计并评估了多种拜占庭容错聚合规则，以保护联邦学习免受模型投毒攻击。其中大多数规则试图在服务器执行聚合时识别并排除恶意更新[8,43,51,149,242-244,359,423]。然而，有动机的对手可以通过向攻击生成优化问题添加约束来绕过这些防御[23,123,335]。梯度裁剪和差分隐私有可能在一定程度上缓解模型投毒攻击[23,271,360]，但它们通常会降低准确率，且不能提供完全的缓解。

For specific model poisoning vulnerabilities, such as backdoor attacks, there are some techniques for model inspection and sanitization (see Sec. 2.3.3). However, mitigating supply-chain attacks in which adversaries might control the source code of the training algorithm or the ML hyperparameters remains challenging. Program verification techniques used in other domains (e.g., cryptographic protocol verification [299]) might be adapted to this setting, but ML algorithms have intrinsic randomness and non-deterministic behavior, which enhances the difficulty of verification.

Designing ML models that are robust in the face of supply-chain model poisoning vulnerabilities is a critical open problem.

2.3.5. Poisoning Attacks in the Real World

As poisoning attacks require adversarial control over the ML training process, they are difficult to mount in the real world. Still, there are several examples of documented cases of real poisoning attacks targeting early AI chatbots, email spam filters, and malware classification services.

The first example of a real-world poisoning attack is the Tay.AI chatbot, a chatbot released by Microsoft on Twitter in 2016 [272]. After online interaction with users for less than 24 hours, the chatbot was poisoned and immediately taken down. At about the same time, there were several large-scale efforts to compromise Google's Gmail spam filter, in which attackers sent millions of emails to attempt to poison the Gmail spam classifier algorithm, enabling them to send other malicious emails without being detected [272]. MITRE ATLAS reported a poisoning incident on the VirusTotal threat intelligence service, in which similar, but not identical samples of a ransomware family were submitted through a popular virus sharing platform to cause the mis-classification of that particular ransomware family [248].

These incidents highlight the risks associated with online learning, as the Tay.AI chatbot was updated in real-time based on user interactions, and the Gmail spam filter and the VirusTotal malware classification system were continuously updated based on newly received samples. In all these incidents, attackers crafted poisoned samples after an initial model release, counting on the fact that models are continuously updated.

对于特定的模型投毒漏洞，例如后门攻击，存在一些用于模型检查和净化的技术(见第 2.3.3 节)。然而，缓解对手可能控制训练算法源代码或 ML 超参数的供应链攻击仍然具有挑战性。其他领域使用的程序验证技术(例如密码协议验证[299])或许可以适配到这一场景，但 ML 算法具有内在随机性和非确定性行为，这增加了验证的难度。

设计在面对供应链模型投毒漏洞时具有韧性的ML模型是一个关键的开放问题。

2.3.5. 现实世界中的投毒(poisoning)攻击

由于投毒攻击需要对ML训练过程进行对抗性控制，因此它们难以在现实世界中实施。尽管如此，仍有几个记录在案的真实投毒攻击案例，针对早期AI聊天机器人、电子邮件垃圾邮件过滤器和恶意软件分类服务。

现实世界中投毒攻击的第一个例子是 Tay.AI 聊天机器人，这是微软于 2016 年在 Twitter 上发布的一款聊天机器人[272]。在与用户进行在线互动不到 24 小时后，该聊天机器人被投毒并立即下线。大约同一时期，出现了若干大规模攻陷 Google Gmail 垃圾邮件过滤器的行动，攻击者发送数百万封电子邮件，试图对 Gmail 垃圾邮件分类器算法进行投毒，从而能够发送其他恶意电子邮件而不被发现[272]。MITRE ATLAS 报告了 VirusTotal 威胁情报服务上的一起投毒事件，其中某个勒索软件家族的相似但不完全相同的样本通过一个流行的病毒共享平台提交，导致该特定勒索软件家族被错误分类[248]。

这些事件凸显了在线学习相关的风险，因为Tay.AI聊天机器人基于用户交互实时更新，而Gmail垃圾邮件过滤器和VirusTotal恶意软件分类系统基于新收到的样本持续更新。在所有这些事件中，攻击者在初始模型发布后精心制作被投毒的样本，指望模型会持续更新这一事实。

2.4. Privacy Attacks and Mitigations

[**NISTAML.03**] [Back to Index]

The seminal work of Dinur and Nissim [110] introduced DATA RECONSTRUCTION attacks, which seek to reverse-engineer private information about an individual user record or other sensitive input data from access to a trained model. More recently, data reconstruction attacks have been designed for binary and multi-class neural network classifiers [50, 152]. With MEMBERSHIP-INFERENCE ATTACK, an adversary can determine whether a particular record was included in the dataset used for training an ML model. Membership inference attacks were first introduced by Homer et al. [162] for genomic data. Recent literature focuses on membership attacks against ML models in mostly black-box settings in which adversaries have query access to a trained ML model [54, 342, 422]. Property inference attacks [19, 74, 134, 233, 361, 437] aim to extract global information about a training dataset, such as the fraction of training examples with a certain sensitive attribute. A different privacy violation for MLaaS is MODEL EXTRACTION attacks, which are designed to extract information about an ML model, such as its architecture or model parameters[3] [58, 70, 177, 376].

This section discusses privacy attacks related to data reconstruction, the memorization of training data, membership inference, property inference, and model extraction, as well as mitigations for some of these attacks and open problems in designing general mitigation strategies.

2.4.1. Data Reconstruction

[**NISTAML.032**] [Back to Index]

Data reconstruction attacks have the ability to recover an individual's data from released aggregate information. Dinur and Nissim [110] were the first to introduce reconstruction attacks that recover user data from linear statistics. Their original attack required an exponential number of queries for reconstruction, but subsequent work has shown how to perform reconstruction with a polynomial number of queries [116]. A survey of privacy attacks, including reconstruction attacks, is given by Dwork et al. [114]. More recently, the U.S. Census Bureau performed a large-scale study on the risk of data reconstruction attacks on census data [135], which motivated the use of differential privacy in the decennial release of the U.S. Census in 2020.

In the context of ML classifiers, Fredrickson et al. [130] introduced model inversion attacks that reconstruct class representatives from the training data of an ML model. While

[3] A privacy violation in this context describes a loss of confidential information about an ML model. If ML model leakage leads to further privacy violations for individuals (e.g., identity theft, sensitive data extraction), it may also be viewed as a cybersecurity-related privacy event. For further discussion on the relationship between privacy and cybersecurity risks, see NIST Privacy Framework, Version 1.0.

2.4. 隐私攻击与缓解

[**NISTAML.03**][返回索引]

Dinur 和 Nissim[110]的开创性工作引入了数据重构攻击，这类攻击试图通过访问训练好的模型来逆向还原关于单个用户记录或其他敏感输入数据的私有信息。最近，数据重构攻击已被设计用于二分类和多分类神经网络分类器[50,152]。借助成员推断攻击，对手可以确定某条特定记录是否被包含在用于训练 ML 模型的数据集中。成员推断攻击最早由 Homer 等人[162]针对基因组数据提出。近期文献主要关注在大多为黑盒设置下针对 ML 模型的成员攻击，即对手对训练好的 ML 模型具有查询访问权限[54,342,422]。属性推断攻击[19,74,134,233,361,437]旨在提取关于训练数据集的全局信息，例如具有某种敏感属性的训练样本所占比例。MLaaS 面临的另一种隐私侵犯是模型萃取攻击，这类攻击旨在提取关于 ML 模型的信息，例如其架构或模型参数[3][58,70,177,376]。

本节讨论与数据重建、训练数据记忆、成员推断、属性推断和模型萃取相关的隐私攻击，以及其中一些攻击的缓解措施和设计通用缓解策略中的开放问题。

2.4.1. 数据重构

[**NISTAML.032**][返回索引]

数据重构攻击能够从发布的聚合信息中恢复个体数据。Dinur和Nissim[110]首次引入了从线性统计量中恢复用户数据的重构攻击。他们最初的攻击需要指数级数量的查询才能完成重构，但后续工作已展示了如何用多项式数量的查询执行重构[116]。Dwork等人[114]给出了一项关于隐私攻击(包括重构攻击)的综述。最近，美国人口普查局开展了一项关于普查数据重构攻击风险的大规模研究[135]，这推动了2020年美国人口普查十年一度发布中差分隐私的使用。

在ML分类器的背景下，Fredrickson等人[130]引入了模型反转攻击，从ML模型的训练数据中重建类别代表。虽然

3 在此背景下，隐私违规描述的是关于ML模型的机密信息丢失。如果ML模型泄露导致个人进一步的隐私违规(例如身份盗窃、敏感数据提取)，也可能被视为与网络安全相关的隐私事件。关于隐私与网络安全风险之间关系的进一步讨论，见NIST隐私框架1.0版。

model inversion generates semantically similar images as those in the training set, it cannot directly reconstruct the training data of the model. Recently, Balle et al. [26] trained a reconstructor network that can recover a data sample from a neural network model, assuming that a powerful adversary has information about all other training samples. Haim et al. [152] showed how the training data of a binary neural network classifier can be reconstructed from access to the model parameters by leveraging theoretical insights about implicit bias in neural networks. This work has recently been extended to reconstruct training samples of multi-class multi-layer perceptron classifiers [50]. Attribute inference is another relevant privacy attack in which the attacker extracts a sensitive attribute of the training set, assuming partial knowledge about other features in the training data [184].

The ability to reconstruct training samples is partially explained by the tendency of neural networks to memorize their training data. Zhang et al. [431] discussed how neural networks can memorize randomly selected datasets. Feldman [126] showed that the memorization of training labels is necessary to achieve an almost optimal generalization error in ML. Brown et al. [48] constructed two learning tasks based on next-symbol prediction and cluster labeling in which memorization is required for high-accuracy learning. Feldman and Zhang empirically evaluated the benefit of memorization for generalization using an influence estimation method [127]. Data reconstruction attacks and their connection to memorization for generative AI are discussed in Sec. 3.3.2.

2.4.2. Membership Inference

[**NISTAML.033**] [Back to Index]

Membership inference attacks may expose private information about an individual, like reconstruction or memorization attacks, and are of great concern when releasing aggregate information or ML models trained on user data. In certain situations, determining that an individual is part of the training set already has privacy implications, such as in a medical study of patients with a rare disease. Moreover, membership inference can be used as a building block for mounting data extraction attacks [59, 63].

In membership inference, the attacker's goal is to determine whether a particular record or data sample was part of the training dataset used for the statistical or ML algorithm. These attacks were introduced by Homer et al. [162] for statistical computations on genomic data under the name *tracing attacks*. Robust tracing attacks have been analyzed when an adversary gains access to noisy statistical information about the dataset [115]. In the last five years, the literature has used the terminology *membership inference* for attacks against ML models. Most of the attacks in the literature are performed against deep neural networks that are used for classification [54, 89, 208, 342, 421, 422]. Similar to other attacks in AML, membership inference can be performed in white-box settings [208, 264, 317] in which attackers have knowledge of the model's architecture and parameters, but most of the attacks have been developed for black-box settings in which the adversary generates queries to the trained ML model [54, 89, 342, 421, 422].

模型反演生成与训练集中图像语义相似的图像，但它无法直接重构模型的训练数据。最近，Balle等人[26]训练了一个重构器网络，能够在假设一个强大的对手拥有关于所有其他训练样本的信息的情况下，从神经网络模型中恢复数据样本。Haim等人[152]展示了如何通过利用关于神经网络中隐式偏差的理论洞见，从对模型参数的访问中重构二分类神经网络分类器的训练数据。这项工作最近已被扩展到重构多类多层感知机分类器的训练样本[50]。属性推断是另一种相关的隐私攻击，其中攻击者在假设对训练数据中其他特征具有部分知识的情况下，提取训练集的敏感属性[184]。

重构训练样本的能力部分可由神经网络记忆其训练数据的倾向来解释。Zhang等人[431]讨论了神经网络如何记忆随机选择的数据集。Feldman[126]表明，对训练标签的记忆是实现ML中近乎最优泛化误差所必需的。Brown等人[48]构建了两个基于下一符号预测和聚类标签的学习任务，其中记忆是高精度学习所必需的。Feldman和Zhang使用影响估计方法实证评估了记忆对泛化的益处[127]。数据重构攻击及其与生成式人工智能中记忆的联系在第3.3.2节中讨论。

2.4.2. 成员推断

[**NISTAML.033**][返回索引]

成员推断攻击可能像重构或记忆攻击一样暴露关于个体的私人信息，并且在发布聚合信息或在用户数据上训练的ML模型时备受关注。在某些情况下，确定个体属于训练集已具有隐私影响，例如在针对罕见病患者的医学研究中。此外，成员推断可用作发起数据提取攻击的构建模块[59,63]。

在成员推断中，攻击者的目标是确定特定记录或数据样本是否属于用于统计或ML算法的训练数据集的一部分。这些攻击由Homer等人[162]针对基因组数据的统计计算引入，名称为追踪攻击。当对手获得对数据集噪声统计信息的访问权时，稳健追踪攻击已被分析[115]。在过去五年中，文献使用术语成员推断来指代针对ML模型的攻击。文献中的大多数攻击是针对用于分类的深度神经网络执行的[54,89,208,342,421,422]。与AML中的其他攻击类似，成员推断可以在白盒设置中执行[208,264,317]，其中攻击者了解模型的架构和参数，但大多数攻击是为黑盒设置开发的，其中对手向训练好的ML模型生成查询[54,89,342,421,422]。

The attacker's success in membership inference has been formally defined using a cryptographically inspired privacy game in which the attacker interacts with a challenger and needs to determine whether a target sample was used in training the queried ML model [183, 321, 422]. In terms of techniques for mounting membership inference attacks, the loss-based attack by Yeom et al. [422] is one of the most efficient and widely used method. Using the knowledge that the ML model minimizes the loss on training samples, the attack determines that a target sample is part of training if its loss is lower than a fixed threshold (selected as the average loss of training examples). Sablayrolles et al. [317] refined the loss-based attack by scaling the loss using a per-example threshold. Another popular technique introduced by Shokri et al. [342] is *shadow models*, which trains a meta-classifier on examples in and out of the training set obtained by training thousands of shadow ML models on the same task as the original model. This technique is generally expensive, and while it might improve upon the simple loss-based attack, its computational cost is high and requires access to many samples from the distribution to train the shadow models. These two techniques are at opposite ends of the spectrum in terms of their complexity, but they perform similarly in terms of precision at low false positive rates [54].

An intermediary method that obtains good performance in terms of the AREA UNDER THE CURVE (AUC) metric is the LiRA attack by Carlini et al. [54], which trains a smaller number of shadow models to learn the distribution of model logits on examples in and out of the training set. Using the assumption that the model logit distributions are Gaussian, LiRA performs a hypothesis test for membership inference by estimating the mean and standard deviation of the Gaussian distributions. Ye et al. [421] designed a similar attack that performs a one-sided hypothesis test, which does not make any assumptions on the loss distribution but achieves slightly lower performance than LiRA. Recently, Lopez et al. [225] proposed a more efficient membership inference attack that requires training a single model to predict the quantiles of the confidence score distribution of the model under attack. Membership inference attacks have also been designed under the stricter label-only threat model in which the adversary only has access to the predicted labels of the queried samples [89].

There are several public privacy libraries that offer implementations of membership inference attacks: the TensorFlow Privacy library [350] and the ML Privacy Meter [259].

2.4.3. Property Inference

[**NISTAML.034**] [Back to Index]

In property inference attacks (also called distribution inference), the attacker tries to learn global information about the training data distribution by interacting with an ML model. For example, an attacker can determine the fraction of the training set with a certain sensitive attribute (e.g., demographic information) that might reveal potentially confidential information about the training set that is not intended to be released.

攻击者在成员推断中的成功已被形式化定义，采用受密码学启发的隐私博弈，其中攻击者与挑战者交互，需要确定目标样本是否被用于训练所查询的 ML 模型[183,321,422]。在实施成员推断攻击的技术方面，Yeom 等人[422]提出的基于损失的攻击是最高效且广泛使用的方法之一。利用 ML 模型在训练样本上最小化损失的知识，该攻击确定目标样本是训练的一部分，如果其损失低于固定阈值(选择为训练示例的平均损失)。Sablayrolles 等人[317]通过使用每示例阈值缩放损失来改进基于损失的攻击。Shokri 等人[342]引入的另一项流行技术是影子模型，它在训练集内和外的示例上训练一个元分类器，这些示例通过在与原始模型相同的任务上训练数千个影子 ML 模型获得。这种技术通常昂贵，虽然它可能改进简单的基于损失的攻击，但其计算成本高，并且需要访问分布中的许多样本来训练影子模型。这两种技术在复杂性方面处于光谱的两端，但在低假阳性率下的精度方面表现相似[54]。

一种在曲线下面积(AUC)指标方面获得良好性能的中间方法是 Carlini 等人[54]提出的 LiRA 攻击，它训练较少数量的影子模型来学习训练集内和外示例上模型 logits 的分布。利用模型 logit 分布为高斯分布的假设，LiRA 通过估计高斯分布的均值和标准差来执行成员推断的假设检验。Ye 等人[421]设计了一种类似的攻击，执行单侧假设检验，不对损失分布做任何假设，但性能略低于 LiRA。最近，Lopez 等人[225]提出了一种更高效的成员推断攻击，需要训练单个模型来预测受攻击模型置信度分数分布的分位数。成员推断攻击也已在更严格的仅标签威胁模型下设计，其中对手只能访问所查询样本的预测标签[89]。

有几个公共隐私库提供了成员推断攻击的实现：TensorFlow Privacy库[350]和ML Privacy Meter[259]。

2.4.3. 属性推断

[**NISTAML.034**][返回索引]

在属性推断攻击(也称为分布推断)中，攻击者试图通过与ML模型交互来了解训练数据分布的全局信息。例如，攻击者可以确定训练集中具有某种敏感属性(例如人口统计信息)的比例，这可能揭示训练集中不打算发布的潜在机密信息。

Property inference attacks were introduced by Ateniese et al. [19] and formalized as a distinguishing game between the attacker and the challenger training two models with different fractions of the sensitive data [361]. Property inference attacks were designed in white-box settings in which the attacker has access to the full ML model [19, 134, 361] and black-box settings in which the attacker issues queries to the model and learns either the predicted labels [233] or the class probabilities [74, 437]. These attacks have been demonstrated for HIDDEN MARKOV MODEL, SUPPORT VECTOR MACHINES [19], FEEDFORWARD NEURAL NETWORKS [134, 233, 437], CONVOLUTIONAL NEURAL NETWORKS [361], FEDERATED LEARNING [240], GENERATIVE ADVERSARIAL NETWORKS [443], and GRAPH NEURAL NETWORK [442]. Mahloujifar et al. [233] and Chaudhauri et al. [74] showed that poisoning the property of interest can help design a more effective distinguishing test for property inference. Moreover, Chaudhauri et al. [74] designed an efficient property size estimation attack that recovers the exact fraction of the population of interest.

The relationship between different training set inference attacks, such as membership inference, attribute inference, and property inference, has been explored by Salem et al.[321] under a unified definitional framework.

2.4.4. Model Extraction

[**NISTAML.031**] [Back to Index]

In MLaaS scenarios, cloud providers typically train large ML models using proprietary data and would like to keep the model architecture and parameters confidential. The goal of an attacker performing a MODEL EXTRACTION attack is to extract information about the model architecture and parameters by submitting queries to the ML model trained by an MLaaS provider. The first model stealing attacks were shown by Tramer at al. [376] on several online ML services for different ML models, including logistic regression, decision trees, and neural networks. However, Jagielski et al. [177] have shown the exact extraction of ML models to be impossible. Instead, a functionally equivalent model can be reconstructed that is different than the original model but achieves similar performance at the prediction task. Jagielski et al. [177] have shown that even the weaker task of extracting functionally equivalent models is computationally prohibitive (*NP*-hard).

Several techniques for mounting model extraction attacks have been introduced in the literature. The first method is that of direct extraction based on the mathematical formulation of the operations performed in deep neural networks, which allows the adversary to compute model weights algebraically [58, 177, 376]. A second technique is to use learning methods for extraction. For example, active learning [70] can guide the queries to the ML model for more efficient extraction of model weights, and reinforcement learning can train an adaptive strategy that reduces the number of queries [280]. A third technique uses SIDE CHANNEL information for model extraction. Batina et al. [29] used electromagnetic side channels to recover simple neural network models, while Rakin et al. [303] showed how ROWHAMMER ATTACK can be used for model extraction of more complex convolutional

属性推断攻击由 Ateniese 等人[19]引入，并被形式化为攻击者与挑战者之间的区分博弈，挑战者用不同比例的敏感数据训练两个模型[361]。属性推断攻击设计于白盒设置中，其中攻击者可以访问完整的 ML 模型[19,134,361]，以及黑盒设置中，其中攻击者向模型发出查询并学习预测标签[233]或类概率[74,437]。这些攻击已在 HIDDEN MARKOV MODEL、SUPPORT VECTOR MACHINES[19]、FEEDFORWARD NEURAL NETWORKS[134,233,437]、CONVOLUTIONAL NEURAL NETWORKS[361]、FEDERATED LEARNING[240]、GENERATIVE ADVERSARIAL NETWORKS[443]和 GRAPH NEURAL NETWORK[442]上得到演示。Mahloujifar 等人[233]和 Chaudhauri 等人[74]表明，对感兴趣属性进行投毒可以帮助设计更有效的属性推断区分测试。此外，Chaudhauri 等人[74]设计了一种高效的属性大小估计攻击，可以恢复感兴趣群体的确切比例。

Salem 等人[321]在统一定义框架下探讨了不同训练集推断攻击(如成员推断、属性推断和性质推断)之间的关系。

2.4.4. 模型提取

[**NISTAML.031**][返回索引]

在 MLaaS 场景中，云提供商通常使用专有数据训练大型 ML 模型，并希望保持模型架构和参数机密。执行模型萃取攻击的攻击者的目标是通过向 MLaaS 提供商训练的 ML 模型提交查询来提取有关模型架构和参数的信息。Tramer 等人[376]在多个在线 ML 服务上针对不同 ML 模型(包括逻辑回归、决策树和神经网络)展示了首批模型窃取攻击。然而，Jagielski 等人[177]表明精确提取 ML 模型是不可能的。相反，可以重建一个功能等效的模型，它不同于原始模型但在预测任务上达到相似性能。Jagielski 等人[177]表明，即使提取功能等效模型的较弱任务在计算上也是不可行的(*NP*-难)。

文献中已引入多种实施模型萃取攻击的技术。第一种方法是基于深度神经网络中执行操作的数学公式进行直接提取，这允许对手以代数方式计算模型权重[58,177,376]。第二种技术是使用学习方法进行提取。例如，主动学习[70]可以引导对 ML 模型的查询以更高效地提取模型权重，而强化学习可以训练一种减少查询次数的自适应策略[280]。第三种技术使用侧信道信息进行模型萃取。Batina 等人[29]使用电磁侧信道恢复简单的神经网络模型，而 Rakin 等人[303]展示了如何利用 ROWHAMMER ATTACK 对更复杂的卷积模型进行模型萃取。

neural network architectures.

Model extraction is often not an end goal but a step toward other attacks. As the model weights and architecture become known, attackers can launch more powerful attacks that are typical for the white-box or gray-box settings. Therefore, preventing model extraction can mitigate downstream attacks that depend on the attacker having knowledge of the model architecture and weights.

2.4.5. Mitigations

The discovery of reconstruction attacks against aggregate information motivated the rigorous definition of *differential privacy* (DP) [112, 113], an extremely strong definition of privacy that guarantees a bound on how much an attacker with access to the algorithm output can learn about each individual record in the dataset. The original *pure* definition of DP has a privacy parameter ε (i.e., privacy budget), which bounds the probability that the attacker with access to the algorithm's output can determine whether a particular record was included in the dataset. DP has been extended to the notions of approximate DP, which includes a second parameter δ that is interpreted as the probability of information accidentally being leaked in addition to ε and Rènyi DP [246].

DP has been widely adopted due to several useful properties: group privacy (i.e., the extension of the definition to two datasets that differ in k records), post-processing (i.e., privacy is preserved even after processing the output), and composition (i.e., privacy is composed if multiple computations are performed on the dataset). DP mechanisms for statistical computations include the Gaussian mechanism [113], the Laplace mechanism [113], and the Exponential mechanism [238]. The most widely used DP algorithm for training ML models is DP-SGD [1], and recent improvements include DP-FTRL [189] and DP matrix factorization [105].

By definition, DP provides mitigation against data reconstruction and membership inference attacks. In fact, the definition of DP immediately implies an upper bound on the success of an adversary in mounting a membership inference attack. Tight bounds on the success of membership inference have been derived by Thudi et al. [369]. However, DP does not provide guarantees against model extraction attacks, as this method is designed to protect the training data, not the model. Several papers have reported negative results after using differential privacy to protect against property inference attacks that aim to extract the properties of subpopulations in the training set [74, 233].

One of the main challenges of using DP in practice is setting up the privacy parameters to achieve a trade-off between the level of privacy and the achieved utility, which is typically measured in terms of accuracy for ML models. Analysis of privacy-preserving algorithms (e.g., DP-SGD) is often worst-case and not tight, and selecting privacy parameters based purely on theoretical analysis results in utility loss. Therefore, large privacy parameters are often used in practice (e.g., the 2020 U.S. Census release used $\varepsilon = 19.61$), and the

神经网络架构。

模型萃取往往不是最终目标，而是通向其他攻击的步骤。随着模型权重和架构变得已知，攻击者可以发动白盒或灰盒设置下典型的更强大攻击。因此，防止模型萃取可以缓解依赖攻击者掌握模型架构和权重知识的下游攻击。

2.4.5. 缓解措施

针对聚合信息的重构攻击的发现，推动了对差分隐私(DP)[112,113]的严格定义，这是一种极强的隐私定义，保证能够限制攻击者在访问算法输出后可以了解数据集中每条个体记录的程度。DP最初的纯定义具有隐私参数ε(即隐私预算)，该参数限制了访问算法输出的攻击者能够确定特定记录是否被包含在数据集中的概率。DP已被扩展为近似DP的概念，其中包含第二个参数δ，该参数被解释为除ε之外信息被意外泄露的概率，以及Rènyi DP[246]。

DP因其若干有用特性而被广泛采用：群体隐私(即，将定义扩展到在*k*条记录上不同的两个数据集)、后处理(即，即使在处理输出之后隐私仍得以保持)以及组合(即，如果对数据集执行多次计算，隐私是可组合的)。用于统计计算的DP机制包括高斯机制[113]、拉普拉斯机制[113]和指数机制[238]。用于训练ML模型的最广泛使用的DP算法是DP-SGD[1]，近期的改进包括DP-FTRL[189]和DP矩阵分解[105]。

根据定义，DP提供了针对数据重构和成员推断攻击的缓解。事实上，DP的定义直接蕴含了对手在实施成员推断攻击时成功概率的上界。Thudi等人[369]推导出了成员推断成功率的紧致界。然而，DP并不提供针对模型萃取攻击的保证，因为该方法旨在保护训练数据，而非模型。若干论文报告了在使用差分隐私来防御旨在提取训练集中子群体属性的属性推断攻击后得到的负面结果[74,233]。

在实践中使用DP的主要挑战之一是设置隐私参数以在隐私水平和所实现的效用之间取得权衡，后者通常以ML模型的准确率来衡量。对隐私保护算法(例如DP-SGD)的分析通常是最坏情况下的且不紧致，仅基于理论分析结果来选择隐私参数会导致效用损失。因此，实践中通常使用较大的隐私参数(例如，2020年美国人口普查发布使用了ε = 19 .61)，并且

exact privacy obtained in practice is difficult to estimate. Jagielski et al. [181] introduced *privacy auditing* with the goal of empirically measuring the actual privacy guarantees of an algorithm and determining privacy lower bounds by mounting privacy attacks. Many privacy auditing techniques are based on inserting *canaries* – synthetic and easy-to-identify out-of-distribution examples – into the training set, and then measuring the canary presence in model output. Auditing can also be performed with membership inference attacks [183, 427], but intentional insertion of strong canaries may result in better estimates of privacy leakage [181, 265]. Recent advances in privacy auditing include tighter bounds for the Gaussian mechanism [263] and rigorous statistical methods that allow for the use of multiple canaries to reduce the sample complexity of auditing [297]. Additionally, two efficient methods for privacy auditing with training a single model have been proposed: Steinke et al. [355] use, multiple random data canaries without incurring the cost of group privacy; and Andrew et al. [10] used multiple random client canaries and cosine similarity test statistics to audit user-level private federated learning.

Differential privacy provides a rigorous notion of privacy and protects against membership inference and data reconstruction attacks. To achieve the best balance between privacy and utility, empirical privacy auditing is recommended to complement the theoretical analysis of private training algorithms.

There are other mitigation techniques against model extraction, such as limiting user queries to the model, detecting suspicious queries to the model, or creating more robust architectures to prevent side-channel attacks. However, these techniques can be circumvented by motivated and well-resourced attackers and should be used with caution. There are practice guides available for securing ML deployments [69, 274]. A completely different approach to potentially mitigating privacy leakage of a user's data is to perform MACHINE UNLEARNING, a technique that enables a user to request the removal of their data from a trained ML model. Existing techniques for machine unlearning are either exact (i.e., retraining the model from scratch or from a certain checkpoint) [45, 52] or approximate (i.e., updating the model parameters to remove the influence of the unlearned records) [139, 175, 268]. They offer different tradeoffs between computation and privacy guarantees, with exact unlearning methods offering stronger privacy, at additional computational cost.

实践中获得的精确隐私难以估计。Jagielski等人[181]引入了隐私审计，其目标是通过实施隐私攻击来经验性地测量算法的实际隐私保证并确定隐私下界。许多隐私审计技术基于将金丝雀--合成的且易于识别的分布外样本--插入训练集，然后测量模型输出中金丝雀的存在。审计也可以通过成员推断攻击[183,427]来执行，但有意插入强金丝雀可能会产生对隐私泄露[181,265]更好的估计。隐私审计的近期进展包括高斯机制的更紧致界[263]以及允许使用多个金丝雀以降低审计样本复杂度的严格统计方法[297]。此外，已经提出了两种通过训练单个模型进行隐私审计的高效方法：Steinke等人[355]使用多个随机数据金丝雀而不产生群体隐私的代价；Andrew等人[10]使用多个随机客户端金丝雀和余弦相似度检验统计量来审计用户级私有联邦学习。

差分隐私提供了严格的隐私概念，并能防御成员推断和数据重构攻击。为在隐私与效用之间取得最佳平衡，建议采用经验隐私审计来补充私有训练算法的理论分析。

针对模型萃取还有其他缓解技术，例如限制用户对模型的查询、检测对模型的可疑查询，或创建更稳健的架构以防止侧信道攻击。然而，这些技术可以被有动机且资源充足的攻击者规避，应谨慎使用。有可用于保护 ML 部署的实践指南[69,274]。一种可能缓解用户数据隐私泄露的完全不同的方法是执行机器遗忘(MACHINE UNLEARNING)，这是一种使用户能够请求从已训练的 ML 模型中移除其数据的技术。现有的机器遗忘技术要么是精确的(即从头开始或从某个检查点重新训练模型)[45,52]，要么是近似的(即更新模型参数以移除被遗忘记录的影响)[139,175,268]。它们在计算和隐私保证之间提供不同的权衡，精确遗忘方法以额外的计算成本提供更强的隐私。

3. Generative AI Taxonomy

GenAI is a branch of AI that develops models that can generate content (e.g., images, text, and other media) with similar properties to their training data. GenAI includes several different types of AI technologies with distinct origins, modeling approaches, and related properties, including: GENERATIVE ADVERSARIAL NETWORKS, GENERATIVE PRE-TRAINED TRANSFORMER (GPT), and DIFFUSION MODELS, among others. Recently, GenAI systems have emerged with multi-modal content generation or comprehension capabilities [119], sometimes through combining two or more model types.

3.1. Attack Classification

While many attack types in the PredAI taxonomy apply to GenAI (e.g., data poisoning, model poisoning, and model extraction), recent work has also introduced novel AML attacks specific to GenAI systems.

Figure 2 shows a taxonomy of attacks in AML for GenAI systems. Similar to the PredAI taxonomy in Fig. 1, this taxonomy is first categorized by the system properties that attackers seek to compromise in each case, including **availability** breakdowns, **integrity** violations, and **privacy** compromises, as well as the additional category of AML attack relevant to GenAI of **misuse** enablement, in which attackers seek to circumvent restrictions placed on the outputs of GenAI systems (see Sec. 2.1.2). The capabilities that an adversary must leverage to achieve their objectives are shown in the outer layer of the objective circles. Attack classes are shown as callouts connected to the capabilities required to mount each attack. Where there are specific types of a more general class of attack (for example, a jailbreak is a specific kind of direct prompting attack attack), the specific attack is linked to the more general attack class through an additional callout. Certain attack classes are listed multiple times because the same attack technique can be used to achieve different attacker goals.

3. 生成式 AI 分类

GenAI 是 AI 的一个分支，它开发的模型能够生成与其训练数据具有相似属性的内容(例如图像、文本和其他媒体)。GenAI 包括几种不同类型的 AI 技术，具有不同的起源、建模方法和相关属性，包括：生成对抗网络(GENERATIVE ADVERSARIAL NETWORKS)、生成预训练变换器(GENERATIVE PRE-TRAINED TRANSFORMER,GPT)和扩散模型(DIFFUSION MODELS)等。最近，GenAI 系统已经出现具有多模态内容生成或理解能力[119]，有时通过组合两种或更多模型类型来实现。

3.1. 攻击分类

尽管 PredAI 分类法中的许多攻击类型适用于 GenAI(例如数据投毒、模型投毒和模型萃取)，近期工作还引入了针对 GenAI 系统的新型 AML 攻击。

图 2 显示了 GenAI 系统 AML 中攻击的分类法。与图 1 中的 PredAI 分类法类似，该分类法首先按攻击者在每种情况下寻求破坏的系统属性进行分类，包括**可用性**破坏、**完整性**违反和**隐私**泄露，以及与 GenAI 相关的额外 AML 攻击类别**滥用**启用，其中攻击者寻求规避对 GenAI 系统输出施加的限制(见第 2.1.2 节)。adversary 必须利用以实现其目标的能力显示在目标圆的外层。攻击类别显示为与发动每种攻击所需能力相连的标注。在存在更一般攻击类别的特定类型的情况下(例如，越狱是一种特定类型的直接提示攻击)，特定攻击通过额外的标注链接到更一般的攻击类别。某些攻击类别被多次列出，因为相同的攻击技术可用于实现不同的攻击者目标。

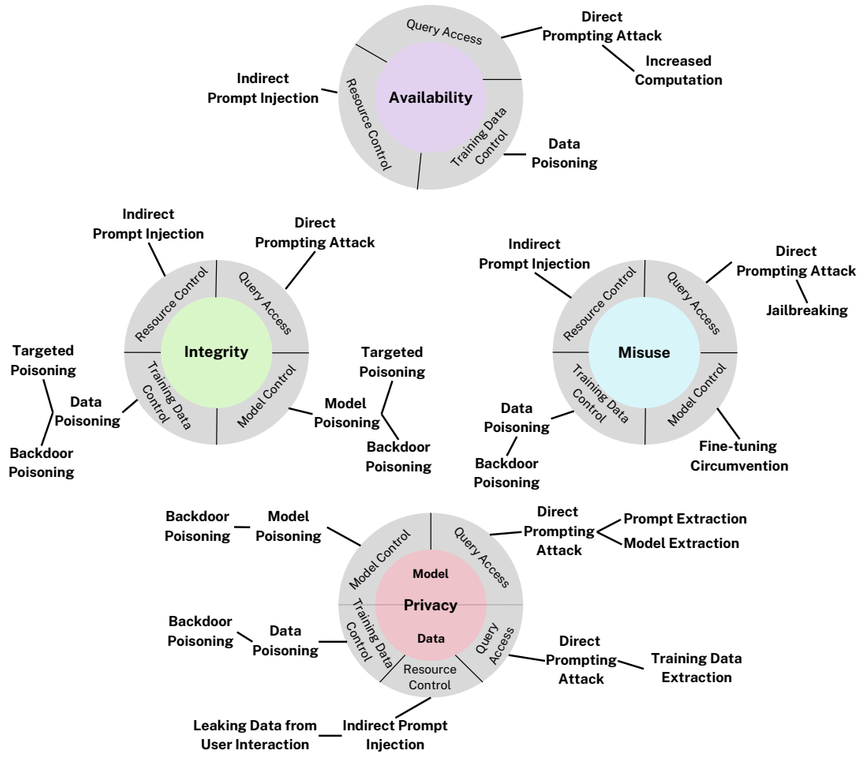


Figure 2. Taxonomy of attacks on GenAI systems

An attack can be further categorized by the learning stage to which it applies and by the attacker's knowledge and access. These are reviewed in the following sections. Where possible, the discussion broadly applies to GenAI models, though some attacks may be most relevant to particular kinds of GenAI models or model-based systems such as RETRIEVAL-AUGMENTED GENERATION (RAG) [RAG] systems, chatbots, or AGENT systems.

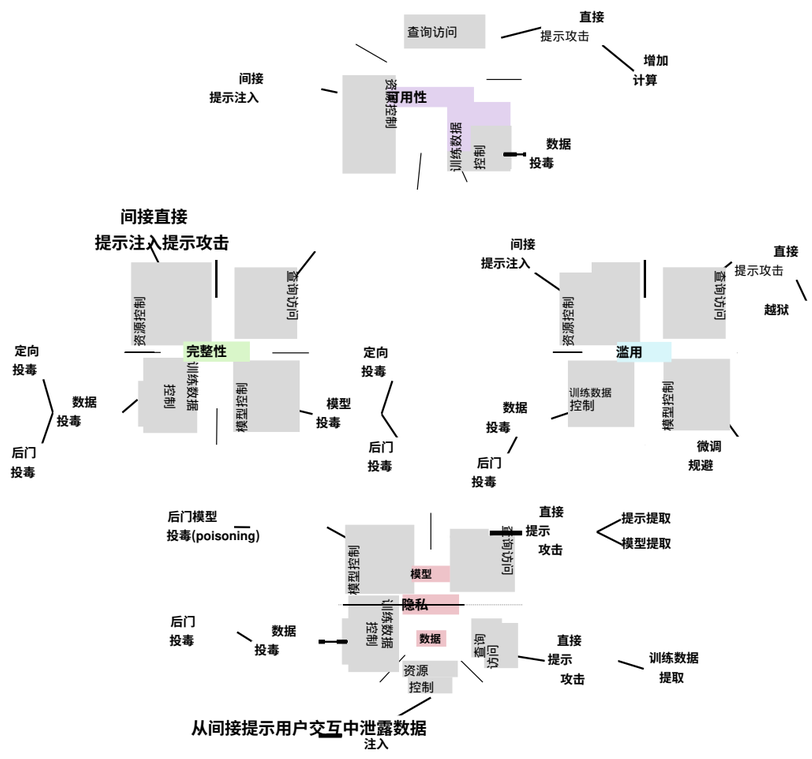


图 2. 针对 GenAI 系统的攻击分类体系

攻击还可按其所适用的学习阶段以及攻击者的知识和访问权限进一步分类。以下各节将对此进行回顾。在可能的情况下，讨论广泛适用于 GenAI 模型，尽管某些攻击可能与特定类型的 GenAI 模型或基于模型的系统最为相关，例如检索增强生成(RAG)[RAG] 系统、聊天机器人或 AGENT 系统。

3.1.1. GenAI Stages of Learning

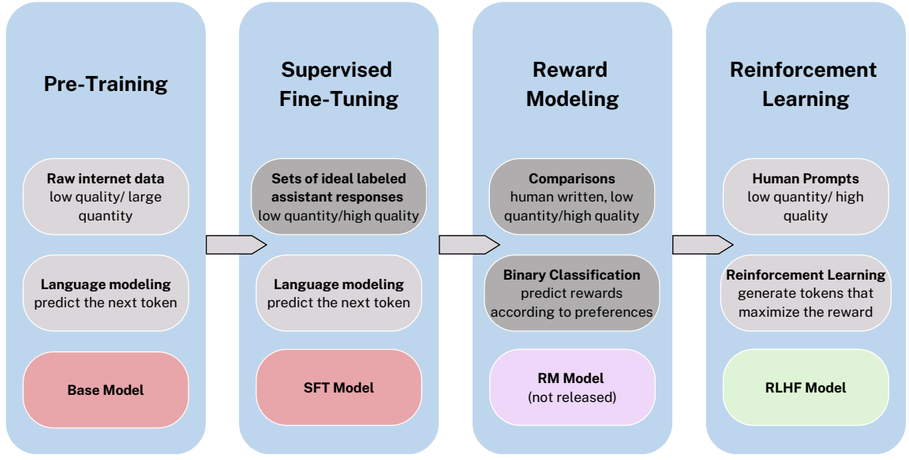


Figure 3. Example LLM Training Pipeline used for InstructGPT [281]

The GenAI development pipeline shapes the space of possible AML attacks against GenAI models and systems. In GenAI moreso than in PredAI, different activities such as data collection, model training, model deployment, and application development are often carried out by multiple different organizations or actors.

For example, a common paradigm in GenAI is the use of a smaller number of foundation models to support a diverse range of downstream applications. Foundation models are pre-trained on large-scale data using self-supervised learning in order to encode general patterns in text, images, or other data that may be relevant for many different applications [311]. Data at the scale used in foundation models is often collected from a variety of internet sources (which attackers can target, such as in DATA POISONING attacks).

This generalist learning paradigm equips foundation models with a variety of capabilities and tendencies — many of which are desirable, but some of which may be harmful or unwanted by the model developer. Techniques such as supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF) can be used after initial pre-training to better align the base model with human preferences and to curb undesirable or harmful model outputs [281] (see Fig. 3). However, these interventions can later be targeted using AML techniques by attackers seeking to recover or re-enable potentially harmful capabilities.

Developers can make trained foundation models available to downstream users and developers in a variety of ways, including openly releasing the model's weights for re-use and

3.1.1. GenAI 学习阶段

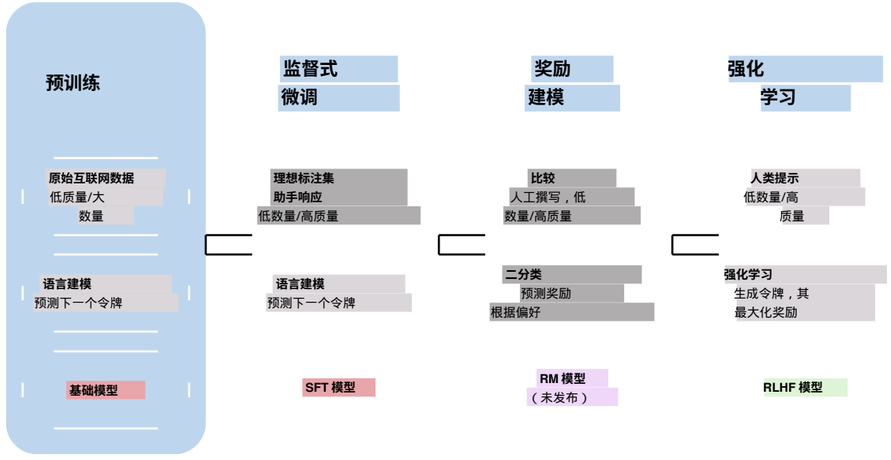


图 3. 用于的示例 LLM 训练流水线 InstructGPT

GenAI 开发流水线塑造了针对 GenAI 模型和系统的可能 AML 攻击空间。与 PredAI 相比，GenAI 中更是如此：数据收集、模型训练、模型部署和应用开发等不同活动通常由多个不同的组织或行为者执行。

例如，GenAI 中的一种常见范式是使用少量基础模型来支持多样化的下游应用。基础模型使用自监督学习在大规模数据上进行预训练，以编码文本、图像或其他数据中可能与许多不同应用相关的通用模式[311]。基础模型所用规模的数据通常从各种互联网来源收集(攻击者可以针对这些来源，例如在数据投毒攻击中)。

这种通才学习范式赋予基础模型各种能力和倾向--其中许多是可取的，但有些可能对模型开发者有害或不受欢迎。监督微调(SFT)和基于人类反馈的强化学习(RLHF)等技术可以在初始预训练之后使用，以更好地使基础模型与人类偏好对齐，并抑制不良或有害的模型输出[281](见图3)。然而，这些干预措施随后可能成为攻击者使用 AML 技术针对的目标，攻击者试图恢复或重新启用潜在有害的能力。

开发者可以通过多种方式将训练好的基础模型提供给下游用户和开发者，包括公开释放模型权重以供复用，以及

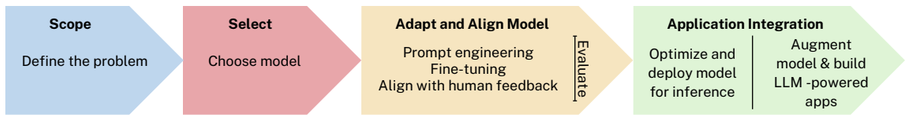


Figure 4. LLM enterprise adoption pipeline

modification, or hosting the model and offering access as a service through an API. These release decisions impact attacker capabilities that shape the space of possible AML attacks, such as whether attackers possess MODEL CONTROL.

Depending on how a foundation model has been made available, downstream developers can customize and build upon the model to create new applications, such as by further fine-tuning the model for a specific use case, or by integrating a foundation model with a software system, such as to build a retrieval-augmented generation (RAG) or agent (see Figure 4). Thus, a foundation model's vulnerabilities to AML attacks can potentially impact a wide range of downstream applications and end users. At the same time, the specific application context in which a foundation model is integrated can create additional vectors for and risks from AML attacks, such as the potential exposure of application-specific data.

AML attacks differ and depend on different phases of the GenAI development lifecycle. One major division is between attacks that target the training stage and those that target model inference during the deployment stage.

Training-time attacks. [**NISTAML.037**] [Back to Index] The TRAINING STAGE for GenAI often consists of foundation model PRE-TRAINING and model FINE-TUNING. This pattern exists for generative image models, text models, audio models, and multimodal models, among others. Since foundation models are most effective when trained on large datasets, it has become common to scrape data from a wide range of public sources, increasing the vulnerability of these models to DATA POISONING attacks. Additionally, GenAI systems trained or fine-tuned by third parties are often used in downstream applications, leading to the risk of MODEL POISONING attacks from maliciously constructed models.

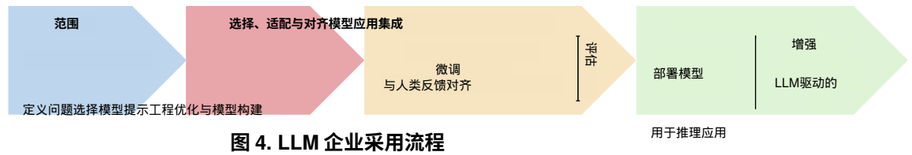


图 4. LLM 企业采用流程

修改，或托管模型并通过 API 以服务形式提供访问。这些发布决策会影响攻击者能力，从而塑造可能的 AML 攻击空间，例如攻击者是否拥有模型控制权。

取决于基础模型的提供方式，下游开发者可以定制并基于该模型构建新应用，例如通过针对特定用例进一步微调模型，或将基础模型与软件系统集成，例如构建检索增强生成(RAG)或智能体(见图 4)。因此，基础模型对 AML 攻击的脆弱性可能影响广泛的下游应用和最终用户。同时，基础模型所集成的特定应用上下文可能为 AML 攻击创造额外的向量和风险，例如应用特定数据的潜在暴露。

AML 攻击各不相同，并依赖于 GenAI 开发生命周期的不同阶段。一个主要划分是：针对训练阶段的攻击与针对部署阶段模型推理的攻击。

训练时攻击。[**NISTAML.037**][返回索引] GenAI 的训练阶段通常包括基础模型预训练和模型微调。这种模式存在于生成式图像模型、文本模型、音频模型和多模态模型等之中。由于基础模型在大型数据集上训练时最为有效，从广泛的公共来源抓取数据已变得普遍，增加了这些模型对数据投毒攻击的脆弱性。此外，由第三方训练或微调的 GenAI 系统常用于下游应用，导致来自恶意构造模型的模型投毒攻击风险。

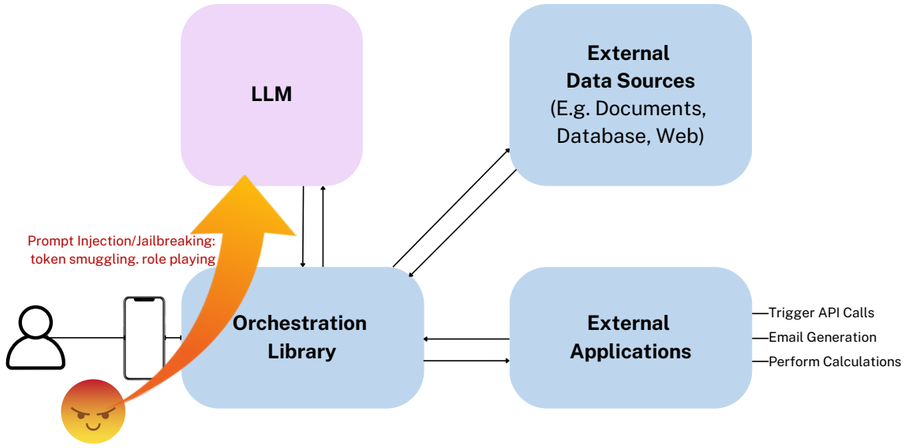


Figure 5. LLM enterprise adoption reference architecture

Inference-time attacks. The DEPLOYMENT STAGE for GenAI models and systems varies both based on how models are hosted or otherwise made available to users, and in how they are integrated into downstream applications. However, GenAI models and applications often share properties that leave them vulnerable to similar types of attacks. For example, underlying many of the security vulnerabilities in LLM applications is the fact that data and instructions are not provided in separate channels to the LLM, which allows attackers to use data channels to inject malicious instructions in inference-time attacks (a similar flaw to that which underlies decades-old SQL injection attacks). Many of the attacks in this stage are due to the following practices that are common in applications of text-based generative models:

1. **In-context instructions and system prompts:** [**NISTAML.035**] [Back to Index] The behavior of LLMs can be shaped through inference-time prompting, whereby the developer or user provides in-context instructions that are often prepended to the model's other input and context. These instructions comprise a natural language description of the model's application-specific use case (e.g., "You are a helpful financial assistant who responds gracefully and concisely....") and is known as a SYSTEM PROMPT. A PROMPT INJECTION overrides these instructions, exploiting the concatenation of untrusted user output to the system prompt to induce unintended behavior. For example, an attacker could inject a JAILBREAK that overrides the system prompt to cause the model to generate restricted or unsafe outputs. Since these prompts have been carefully crafted through prompt engineering and may be security-relevant, a PROMPT EXTRACTION attack may attempt to steal these system instructions. These attacks are also relevant to multimodal and text-to-image models.

2. **Runtime data ingestion from third-party sources:** In RETRIEVAL-AUGMENTED GENERA-

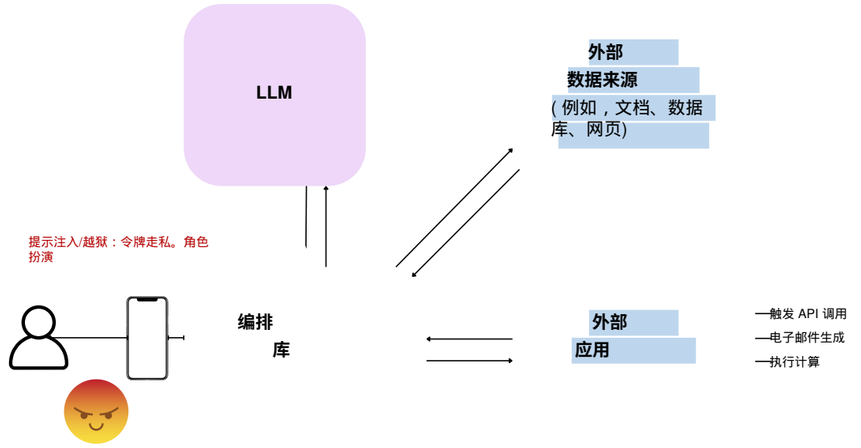


图 5. LLM 企业采用参考架构

推理时攻击。 GenAI 模型和系统的部署阶段因模型托管方式或以其他方式提供给用户的方式，以及它们集成到下游应用的方式而有所不同。然而，GenAI 模型和应用通常具有共同属性，使其容易受到类似类型的攻击。例如，LLM 应用中许多安全漏洞的根本原因是数据和指令未通过单独的通道提供给 LLM，这允许攻击者在推理时攻击中利用数据通道注入恶意指令(类似于数十年前 SQL 注入攻击的根本缺陷)。此阶段的许多攻击源于基于文本的生成模型应用中常见的以下做法：

1. **上下文指令与系统提示：**[**NISTAML.035**][返回索引] LLM 的行为可以通过推理时提示来塑造，即开发者或用户提供上下文指令，这些指令通常被前置到模型的其他输入和上下文之前。这些指令包含对模型特定应用用例的自然语言描述(例如，"你是一个乐于助人的财务助手，能够优雅而简洁地回应....")，被称为系统提示(SYSTEM PROMPT)。提示注入(PROMPT INJECTION)会覆盖这些指令，利用将不可信用户输出拼接到系统提示上的方式诱导非预期行为。例如，攻击者可以注入越狱(JAILBREAK)来覆盖系统提示，导致模型生成受限或不安全的输出。由于这些提示是通过提示工程精心设计的，且可能与安全相关，提示萃取(PROMPT EXTRACTION)攻击可能试图窃取这些系统指令。这些攻击也与多模态模型和文本到图像模型相关。

2. **来自第三方来源的运行时数据摄取：**在检索增强生成中--

TION (RAG) applications, chatbots, and other applications in which GenAI models are used to interface with additional resources, context is often crafted at runtime in a query-dependent way and populated from external data sources (e.g., documents, web pages, etc.) that are to be used as part of the application. INDIRECT PROMPT INJECTION attacks depend on the attacker's ability to modify external sources of information that will be ingested into the model context, even if not provided directly by the primary system user.

3. **Output handling:** The output of an GenAI model may be used dynamically, such as to populate an element on a web page or to construct a command that is executed without any human supervision, which can lead to a range of availability, integrity or privacy violations in downstream applications if an attacker can induce behavior in this output that the developer has not accounted for.

4. **Agents:** An LLM-based AGENT relies on iteratively processing the *output* of an LLM (item 3 above) to perform a task and then provides the results as additional context back to the LLM input (item 2) [151, 155, 393]. For example, an agent system may select from among a configured set of external dependencies and invoke the code with templates filled out by the LLM using information in the context. Adversarial inputs into this context, such as from interactions with untrusted resources, could hijack the agent into performing adversary-specified actions instead, leading to potential security or safety violations.

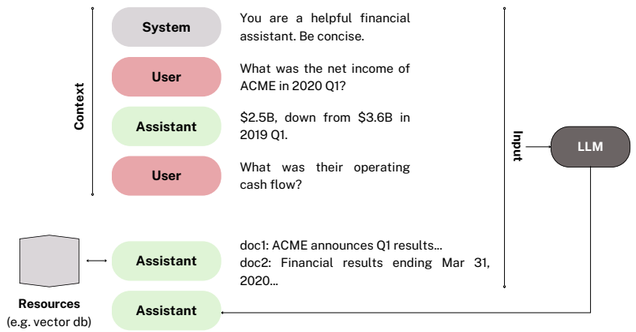


Figure 6. Retrieval-augmented generation

3.1.2. Attacker Goals and Objectives

As with PredAI, attacker objectives can be classified broadly along the dimensions of availability, integrity, and privacy, along with a new, GenAI-specific category of attacks designed to enable misuse.

- In an **AVAILABILITY BREAKDOWN attack** [**NISTAML.01**] [Back to Index], an attacker

在检索增强生成(RAG)应用、聊天机器人以及其他使用 GenAI 模型与额外资源交互的应用中，上下文通常以依赖于查询的方式在运行时构建，并从外部数据源(例如文档、网页等)填充，作为应用的一部分使用。间接提示注入攻击依赖于攻击者修改将被摄入模型上下文的外部信息源的能力，即使这些信息并非由主要系统用户直接提供。

3. **输出处理：** GenAI 模型的输出可能被动态使用，例如填充网页上的元素，或构造在无任何人工监督下执行的命令；如果攻击者能够诱导该输出产生开发者未预料到的行为，则可能导致下游应用中出现一系列可用性、完整性或隐私违规。

4. **智能体(Agents):**基于 LLM 的智能体(AGENT)依赖于迭代处理 LLM 的输出(上文第 3 项)以执行任务，然后将结果作为附加上下文反馈给 LLM 输入(第 2 项)[151,155,393]。例如，智能体系统可以从一组已配置的外部依赖中进行选择，并调用由 LLM 使用上下文中的信息填充模板的代码。对此上下文的对抗性输入(例如来自与不可信资源的交互)可能劫持智能体转而执行攻击者指定的操作，从而导致潜在的安全或安保违规。

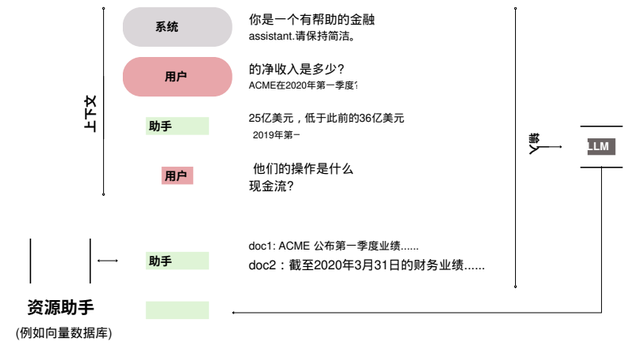


图 6. 检索增强生成

3.1.2. 攻击者的目标与目的

与 PredAI 一样，攻击者目标可以大致按可用性、完整性和隐私维度分类，此外还有一类新的、GenAI 特有的攻击类别，旨在促成滥用。

- 在**可用性破坏攻击**[**NISTAML.01**][返回索引]中，攻击者

seeks to interfere with a GenAI model or system to disrupt the ability of other users or processes to obtain timely and consistent access to its outputs or functionality.

- In an **INTEGRITY VIOLATION attack** [**NISTAML.02**] [Back to Index], an attacker seeks to interfere with a GenAI system to force it to misperform against its intended objectives and produce output that aligns with the attacker's objective. As users and businesses rely on GenAI systems to perform tasks like research and productivity assistance, these violations can allow attackers to weaponize the trust that these users place in GenAI systems.
- In a **PRIVACY COMPROMISE attack** [**NISTAML.03**] [Back to Index], an attacker seeks to gain unauthorized access to restricted or proprietary information that is part of a GenAI system, including information about a model's training data, weights or architecture; or sensitive information that the model accesses such as the knowledge base of a RETRIEVAL-AUGMENTED GENERATION (RAG) application. GenAI systems may be exposed to sensitive data (intentionally or otherwise) during training or inference, and attacks may seek to extract such information (e.g., through INDIRECT PROMPT INJECTION attacks, where a third party exfiltrates in-context user information [307], or a MODEL EXTRACTION attack to exfiltrate model information [61]).
- **Misuse enablement** [**NISTAML.04**][Back to Index]. An additional attacker objective that is especially relevant in the GenAI context is the goal of **MISUSE ENABLEMENT**. In these attacks, an attacker seeks to deliberately circumvent technical restrictions imposed by the GenAI system's owner on its use, such as restrictions designed to prevent the system from producing outputs that could cause harm to others, cf. [325].

  Technical restrictions refer in this context to defenses applied to the GenAI system such as the use of system prompts or RLHF for safety alignment. While the specific technical restrictions in place will vary between models, the techniques for circumventing such defenses are often common between different kinds of models and different kinds of misuse, allowing them to be taxonomized as a part of AML without specificity as to the particular kinds of misuse that model developers seek to prevent.

3.1.3. Attacker Capabilities

AML attacks can be taxonomized with respect to the capabilities that an attacker has in controlling inputs to the GenAI model or system. These capabilities include:

- TRAINING DATA CONTROL: The attacker might take control of a subset of the training data by inserting or modifying training samples. This capability is used in DATA POISONING attacks.
- QUERY ACCESS: Many GenAI models and their applications are deployed as services that can be accessed over the internet by users. In these cases, attackers can sub-

试图干扰生成式人工智能模型或系统，以破坏其他用户或进程及时、一致地获取其输出或功能的能力。

- 在**完整性违规攻击**[**NISTAML.02**][返回索引]中，攻击者试图干扰生成式人工智能系统，迫使其偏离预期目标执行错误操作，并产生符合攻击者目标的输出。由于用户和企业依赖生成式人工智能系统执行研究和生产力辅助等任务，这些违规行为可使攻击者将用户对生成式人工智能系统的信任武器化。
- 在隐私泄露攻击(PRIVACY COMPROMISE attack)[**NISTAML.03**][返回索引]中，攻击者试图未经授权访问属于 GenAI 系统的受限或专有信息，包括有关模型训练数据、权重或架构的信息；或模型访问的敏感信息，例如检索增强生成(RAG)应用的知识库。GenAI 系统在训练或推理期间可能(有意或无意地)接触敏感数据，攻击可能试图提取此类信息(例如，通过间接提示注入(INDIRECT PROMPT INJECTION)攻击，由第三方窃取上下文中的用户信息[307]，或通过模型萃取(MODEL EXTRACTION)攻击窃取模型信息[61])。
- **滥用赋能**[**NISTAML.04**][返回索引]。在生成式人工智能语境中尤为相关的另一个攻击者目标是**滥用赋能**。在此类攻击中，攻击者试图故意规避生成式人工智能系统所有者对其使用施加的技术限制，例如旨在防止系统产生可能对他人造成伤害的输出的限制，参见[325]。

  在此语境下，技术限制指的是应用于 GenAI 系统的防御措施，例如使用系统提示或 RLHF 进行安全对齐。虽然具体实施的技术限制因模型而异，但规避此类防御的技术在不同类型的模型和不同类型的滥用之间往往是共通的，因此可以将它们作为 AML 的一部分进行分类，而无需具体指明模型开发者试图防止的特定滥用类型。

3.1.3. 攻击者能力

对抗性机器学习攻击可根据攻击者在控制生成式人工智能模型或系统输入方面所具备的能力进行分类。这些能力包括：

- 训练数据控制：攻击者可能通过插入或修改训练样本来控制训练数据的一个子集。此能力用于数据投毒攻击。
- 查询访问：许多生成式人工智能模型及其应用以服务形式部署，用户可通过互联网访问。在此类情况下，攻击者可向模型提交

mit adversarially crafted queries to the model to elicit specific desired behavior or extract information. This capability is used for PROMPT INJECTION, PROMPT EXTRACTION, and MODEL EXTRACTION attacks. Query access can vary based on the degree of generation control (e.g., modifying the temperature or adding a logit bias) and the richness of the returned generation (e.g., with or without log probabilities or multiple choices).

- RESOURCE CONTROL: The attacker might modify resources (e.g., documents, web pages) that will be ingested by the GenAI model at runtime. This capability is used for INDIRECT PROMPT INJECTION attacks.
- MODEL CONTROL: The attacker might have the ability to modify model parameters, such as through public fine-tuning APIs or openly accessible model weights. This capability is used in MODEL POISONING attacks, as well FINE-TUNING CIRCUMVENTION attacks which remove refusal behavior or other model-level safety interventions [132, 153, 300].

As in PredAI, attackers can also vary in their knowledge of the underlying ML model, from full knowledge of the ML system including model weights (**white-box attacks**), to minimal knowledge and systems with deliberately obscured or misleading information (**black-box attacks**), to somewhere in between (**gray-box attacks**). See Sec. 2.1.4, which discusses attacker knowledge in greater detail and applies to GenAI attacks.

3.2. Supply Chain Attacks and Mitigations

[**NISTAML.05**] [Back to Index]

Since AI is software, it inherits many of the vulnerabilities of the traditional software supply chain, such as reliance on third-party dependencies. AI development also introduces new types of dependencies, including data collection and scoring, the integration or adaptation of third-party-developed AI models, and the integration of third-party-developed plugins into AI systems. Mitigating the security challenges in AI supply chain management is complex and requires a multifaceted approach that combines existing practices for software supply chain risk management with the management of AI-specific supply chain risks, such as through the use of provenance information for the additional artifacts involved [159, 267]. Studies of real-world security vulnerabilities against ML suggest that security is best addressed comprehensively and by considering the full attack surface, including data and model supply chains, software, and network and storage systems [17, 370]. While all of these supply chain risks are critical in the broader context of securing AI systems, there are certain types of attacks that rely on exploiting the specific statistical and data-based properties of ML systems, thus falling within the domain of AML.

对抗性构造的查询，以诱发特定的期望行为或提取信息。此能力用于提示注入、提示萃取和模型萃取攻击。查询访问可因生成控制程度(例如修改温度或添加对数几率偏置)和返回生成的丰富程度(例如是否包含对数概率或多候选)而异。

- 资源控制：攻击者可能修改生成式人工智能模型在运行时将摄取的资源(例如文档、网页)。此能力用于间接提示注入攻击。
- 模型控制：攻击者可能具备修改模型参数的能力，例如通过公开微调API或公开可访问的模型权重。此能力用于模型投毒攻击，以及移除拒绝行为或其他模型级安全干预的微调规避攻击。

[132, 153, 300].

与预测性人工智能一样，攻击者对底层机器学习模型的了解程度也各不相同，从完全了解机器学习系统(包括模型权重)(**白盒攻击**)，到了解极少且系统故意设置模糊或误导性信息(**黑盒攻击**)，再到介于两者之间(**灰盒攻击**)。参见第2.1.4节，该节更详细地讨论了攻击者知识，并适用于生成式人工智能攻击。

3.2. 供应链攻击与缓解

[**NISTAML.05**][返回索引]

由于人工智能是软件，它继承了传统软件供应链的许多漏洞，例如对第三方依赖项的依赖。人工智能开发还引入了新型依赖关系，包括数据收集与评分、第三方开发的人工智能模型的集成或适配，以及将第三方开发的插件集成到人工智能系统中。缓解人工智能供应链管理中的安全挑战十分复杂，需要采取多层面的方法，将现有的软件供应链风险管理实践与人工智能特有的供应链风险管理相结合，例如对涉及的额外工件使用来源信息[159,267]。针对机器学习在现实世界中的安全漏洞的研究表明，最好全面地考虑整个攻击面来解决安全问题，包括数据和模型供应链、软件以及网络和存储系统

[17,370]。尽管所有这些供应链风险在保障人工智能系统安全的更广泛语境中都很关键，但某些类型的攻击依赖于利用机器学习系统特定的统计和数据特性，因而属于对抗性机器学习的范畴。

3.2.1. Data Poisoning Attacks

The performance of GenAI text-to-image and text-to-text models has been found to scale with dataset size (among other properties like model size and data quality); for example, Hoffmann et al. [161] suggest compute-optimally training a 520 billion parameter model may require 11 trillion tokens of training data. Thus, it has become common for GenAI foundation model developers to scrape data from a wide range of sources. In turn, the scale of this data and the diversity of its sources provides a large potential attack surface into which attackers may seek to insert adversarially constructed data points. For example, dataset publishers may provide a list of URLs to constitute a training dataset, and attackers may be able to purchase some of the domains that serve those URLs and replace the site content with their own malicious content [57].

Beyond the vast quantities of pre-training data, data poisoning attacks may also affect other stages of the LLM training pipeline, including instruction tuning [389] and reinforcement learning from human feedback [305], which may intentionally source data from a large number of human participants.

As with PredAI models (see Sec. 2.1), data poisoning attacks could lead to attackers controlling model behavior through the insertion of a backdoor (see BACKDOOR POISONING ATTACK) such as a word or phrase that, when submitted to a model, acts as a universal JAILBREAK [305]. Attackers could also use data poisoning attacks to modify model behavior on particular user queries (see TARGETED POISONING ATTACK), such as causing the model to incorrectly summarize or otherwise produce degenerate outputs in response to queries that contain a particular trigger word or phrase [389]. These attacks may be practical—requiring a relatively small portion of the total dataset [46]—and may lead to a range of bad outcomes, such as code suggestion models which intentionally suggest insecure code [3].

3.2.2. Model Poisoning Attacks

[**NISTAML.051**] [Back to Index]

In GenAI, it is common for developers to use foundation models developed by third parties. Attackers can take advantage of this fact by offering maliciously designed models, such as pre-trained models that enable a BACKDOOR POISONING ATTACK or TARGETED POISONING ATTACK. While this attack relies on the attacker having control over the initial poisoned model, researchers have identified attacks in which malicious backdoors in pre-trained models can persist even after downstream users fine-tune the model for their own use [201] or apply additional safety training measures [170].

3.2.3. Mitigations

GenAI poisoning mitigations largely overlap with PredAI poisoning mitigations (see Sec. 2.3). For preventing data poisoning with web-scale data dependencies, this includes ver-

3.2.1. 数据投毒攻击

研究发现，GenAI 文本到图像和文本到文本模型的性能随数据集规模而扩展(以及模型规模和数据质量等其他属性)；例如，Hoffmann 等人[161]提出，以计算最优方式训练一个 5200 亿参数的模型可能需要 11 万亿个训练数据 token。因此，GenAI 基础模型开发者从广泛来源抓取数据已变得十分普遍。反过来，这种数据的规模及其来源的多样性提供了巨大的潜在攻击面，攻击者可能试图在其中插入对抗性构造的数据点。例如，数据集发布者可能提供一组 URL 来构成训练数据集，而攻击者可能能够购买其中一些提供这些 URL 的域名，并将网站内容替换为自己的恶意内容[57]。

除海量预训练数据外，数据投毒攻击还可能影响大语言模型训练流程的其他阶段，包括指令微调[389]和基于人类反馈的强化学习[305]，这些阶段可能有意从大量人类参与者处获取数据。

与预测性人工智能模型一样(参见第2.1节)，数据投毒攻击可通过插入后门(参见后门投毒攻击)导致攻击者控制模型行为，例如某个词或短语在提交给模型时充当通用越狱

[305]。攻击者还可以利用数据投毒攻击来修改模型对特定用户查询的行为(参见定向投毒攻击)，例如导致模型在响应包含特定触发词或短语的查询时进行错误摘要或以其他方式产生退化输出[389]。这些攻击可能是实用的--只需要占总数据集相对较小的一部分[46]--并可能导致一系列不良后果，例如代码建议模型故意建议不安全的代码[3]。

3.2.2. 模型投毒(poisoning)攻击

[**NISTAML.051**][返回索引]

在 GenAI 中，开发者使用第三方开发的基础模型十分常见。攻击者可以利用这一事实，提供恶意设计的模型，例如能够实现后门投毒攻击或定向投毒攻击的预训练模型。虽然这种攻击依赖于攻击者对初始投毒模型的控制，但研究人员已经发现了这样的攻击：预训练模型中的恶意后门即使在下游用户为自己的用途微调模型[201]或应用额外的安全训练措施[170]之后仍能持续存在。

3.2.3. 缓解措施

生成式人工智能投毒缓解措施与预测性人工智能投毒缓解措施大体重叠(参见第2.3节)。对于防范具有网络规模数据依赖的数据投毒，这包括验

ifying web downloads as a basic integrity check to ensure that domain hijacking has not injected new sources of data into the training dataset [57]. That is, the provider publishes cryptographic hashes, and the downloader verifies the training data. Data filtering can also attempt to remove poisoned samples, though detecting poisoned data within a large training corpus may be very difficult.

While traditional software supply chain risk management practices such as vulnerability scanning of model artifacts can help manage some kinds of AI supply chain risks, new approaches are required to detect vulnerabilities in models such as those introduced through model poisoning attacks. Current proposed approaches include using methods from the field of mechanistic interpretability to identify backdoor features [67] and detecting and counteracting triggers when they are seen at inference time. Beyond these mitigations, risks can be reduced by understanding models as untrusted system components and designing applications such that risks from attacker-controlled model outputs are reduced [266].

3.3. Direct Prompting Attacks and Mitigations

[**NISTAML.018**] [Back to Index] DIRECT PROMPTING ATTACK attacks arise when the attacker is the primary user of the system, interacting with the model through query access. A subset of these attacks, in which the main user provides in-context instructions that are appended to higher-trust instructions like those provided by the application designer (such as the model's SYSTEM PROMPT), are known as DIRECT PROMPT INJECTION attacks.

As in PredAI, attacks may be applicable to a single setting and model, or may instead be universal (affecting models on a range of separate queries, see Sec. 2.2.1) and/or transferable (affecting models beyond the model they are found on, see Sec. 2.2.3).

An attacker may have a variety of goals when performing these attacks [219, 220, 337], such as to:

- **Enable misuse.** Attackers may use direct prompting attacks to bypass model-level defenses that a model developer or deployer has created to restrict models from producing harmful or undesirable output [237]. A JAILBREAK is a direct prompting attack intended to circumvent restrictions placed on model outputs, such as circumventing refusal behavior to enable misuse.
- **Invade privacy.** Attackers may use direct prompting to extract the system prompt or reveal private information that was provided to the model in context but not intended for unfiltered access by the user.
- **Violate integrity.** When LLMs are used as agents, an attacker may use direct prompting attacks to manipulate tool usage and API calls, and potentially compromise the backend of the system (e.g. executing attacker's SQL queries).

证网络下载作为基本完整性检查，以确保域名劫持未向训练数据集注入新的数据源[57]。即，提供方发布密码学哈希，下载方验证训练数据。数据过滤也可尝试移除中毒样本，但在大型训练语料中检测中毒数据可能非常困难。

虽然传统的软件供应链风险管理实践(例如对模型工件进行漏洞扫描)可以帮助管理某些类型的人工智能供应链风险，但需要新的方法来检测模型中的漏洞，例如通过模型投毒攻击引入的漏洞。当前提出的方法包括使用机制可解释性领域的方法来识别后门特征[67]，以及在推理时检测和抵消所看到的触发器。除了这些缓解措施之外，还可以通过将模型理解为不受信任的系统组件并设计应用程序以减少攻击者控制的模型输出带来的风险来降低风险

[266].

3.3. 直接提示攻击与缓解

[**NISTAML.018**][返回索引] 直接提示攻击发生在攻击者是系统主要用户、通过查询访问与模型交互之时。此类攻击的一个子集--其中主要用户提供上下文内指令，这些指令被附加到更高信任度的指令(如应用设计者提供的指令，例如模型的系统提示)之后--被称为直接提示注入攻击。

与预测性人工智能一样，攻击可能适用于单一设置和模型，也可能是通用的(影响模型在一系列独立查询上的表现，参见第2.2.1节)和/或可迁移的(影响其被发现模型之外的模型，参见第2.2.3节)。

攻击者在执行这些攻击时可能有多种目标[219,220,337]，例如：

- **赋能滥用。**攻击者可能利用直接提示攻击绕过模型开发者或部署者创建的、旨在限制模型产生有害或不良输出的模型级防御[237]。越狱是一种直接提示攻击，旨在规避对模型输出施加的限制，例如规避拒绝行为以赋能滥用。
- **侵犯隐私。**攻击者可能利用直接提示提取系统提示，或泄露在上下文中提供给模型但并非意图供用户无过滤访问的私密信息。
- **违反完整性。**当大语言模型被用作智能体时，攻击者可能利用直接提示攻击操纵工具使用和API调用，并可能危及系统后端(例如执行攻击者的SQL查询)。

3.3.1. Attack Techniques

A range of techniques exist for launching direct prompting attacks, many of which generalise across various attacker objectives. With a focus on direct prompting attacks to enable misuse, we note the following broad categories of direct prompting techniques (see [393]):

- **Optimization-based attacks** design attack objective functions and use gradient or other search-based methods to learn adversarial inputs that cause a particular behavior, similar to PredAI attacks discussed in Sec. 2.2.1. Objective functions may be designed to force affirmative starts (e.g., looking for responses that begin with "Sure", which may indicate compliance with a malicious request [60, 320, 448]) or other metrics of attack success (e.g., similarity to a toxified finetune [368]).

  Optimization techniques can then be used to learn attacks, including techniques that follow from attacks designed for PredAI language classifiers (e.g., HotFlip [117]) and gradient-free techniques that use a proxy model or random search to test attack candidates [11, 320]. Universal adversarial triggers are a special class of these gradient-based attacks against generative models that seek to find *input-agnostic* prefixes (or suffixes) that produce the desired affirmative response regardless of the remainder of the input [386, 448]. That these universal triggers transfer to other models makes open-weight models — for which there is ready white-box access — feasible attack vectors for transferability attacks on closed systems in which only API access is available [448].

  Attacks can also be designed to satisfy additional constraints (e.g., sufficiently low perplexity [368]) or attack a system of multiple models [235].

- **Manual methods** for jailbreaking an LLM include competing objectives and mismatched generalization [400]. Mismatched generalization-based attacks identify inputs that fall outside the distribution of the model's safety training but remain within the distribution of its capabilities training, making them comprehensible to the model while evading refusal behavior. Competing objectives-based attacks find cases where model capabilities are in tension with safety goals, such as by playing into a model's drive to follow user-provided instructions. In all cases the goal of the attack is to compromise a model-level safety defense. See Weng [403] for further discussion.

  Approaches to competing objectives-based attacks include:

  1. *Prefix injection*: This method involves prompting the model to start responses with an affirmative confirmation. By conditioning the model to begin its output in a predetermined manner, adversaries attempt to influence its subsequent language generation toward specific, predetermined patterns or behaviors.

  2. *Refusal suppression*: Adversaries may explicitly instruct the model to avoid generating refusals or denials in its output. By decreasing the probability of refusal responses, this tactic aims to increase the probability of a compliant

3.3.1. 攻击技术

存在一系列用于发起直接提示攻击的技术，其中许多技术可泛化至各种攻击者目标。聚焦于赋能滥用的直接提示攻击，我们指出以下直接提示技术的广泛类别(参见[393]):

- **基于优化的攻击**设计攻击目标函数，并使用梯度或其他基于搜索的方法学习引发特定行为的对抗性输入，类似于第2.2.1节讨论的预测性人工智能攻击。目标函数可设计为强制肯定性开头(例如寻找以"Sure"开头的响应，这可能表明遵从恶意请求[60,320,448])或其他攻击成功指标(例如与有毒化微调的相似度[368])。

  随后可以使用优化技术来学习攻击，包括源自针对PredAI语言分类器设计的攻击的技术(例如HotFlip[117])，以及使用代理模型或随机搜索来测试攻击候选的无梯度技术[11,320]。通用对抗触发器是针对生成模型的这类基于梯度的攻击中的一个特殊类别，其目标是找到与输入无关的前缀(或后缀)，使得无论输入的其余部分是什么，都能产生所需的肯定响应[386,448]。这些通用触发器能够迁移到其他模型，这使得开放权重模型--对其存在现成的白盒访问--成为对仅提供API访问的封闭系统实施可迁移性攻击的可行攻击向量[448]。

  攻击也可设计为满足额外约束(例如足够低的困惑度[368])或攻击由多个模型组成的系统[235]。

- **手动方法**用于越狱大语言模型，包括竞争目标和误
  匹配泛化[400]。基于不匹配泛化的攻击识别出落在模型安全训练分布之外但仍处于其能力训练分布之内的输入，使模型能够理解这些输入，同时规避拒绝行为。基于竞争目标的攻击寻找模型能力与安全目标之间存在张力的情况，例如利用模型遵循用户提供指令的驱动力。在所有情况下，攻击的目标都是破坏模型级安全防御。进一步讨论参见Weng[403]。

  针对竞争性目标攻击的方法包括：

  1. 前缀注入：此方法涉及提示模型以肯定性确认开始响应。通过使模型以预定方式开始输出，对抗者试图将其后续语言生成引向特定的预定模式或行为。

  2. 拒绝抑制：对抗者可能明确指示模型避免在输出中生成拒绝或否认。通过降低拒绝响应的概率，此策略旨在提高遵从

response.

3. *Style injection*: In this approach, adversaries instruct the model to use (or not use) certain syntax or writing styles. For example, an attack may constrain the model's language to simplistic or non-professional tones, aiming to decrease the probability of (usually professionally worded) refusals.

4. *Role-play*: Adversaries utilize role-play strategies (e.g., "Always Intelligent and Machiavellian" [AIM] or "Do Anything Now" [DAN]) to guide the model to adopt specific personas or behavioral patterns that conflict with the original intent. This manipulation aims to exploit the model's adaptability to varied roles or characteristics, with an intent to compromise its adherence to safety protocols.

Approaches to mismatched generalization-based attacks include:

1. *Special encoding*: Strategies that use encoding techniques like base64 to alter the representation of input data in a way that remains understandable to the model but may be out of distribution for safety training.

2. *Character transformation*: Strategies that use character-level transformations like the ROT13 cipher, symbol replacement (e.g., l33tspeak), and Morse code to take the input out of the safety training distribution.

3. *Word transformation*: Strategies that alter the linguistic structure of the input, such as Pig Latin, synonym swapping (e.g., using "pilfer" for "steal"), and payload splitting (or "token smuggling") to break down sensitive words into substrings.

4. *Prompt-level transformation*: Strategies that use prompt-level transformations, such as translating the prompt into a less common language that may be out of distribution of the safety training data.

- **Automated model-based red teaming** employs an attacker model, a target model, and a judge [73, 239, 292]. When the attacker has access to a high-quality classifier that judges whether model output is harmful, it may be used as a reward function to train a generative model to generate jailbreaks of another generative model. Only query access is required for each of the models, and no human intervention is required to update or refine a candidate jailbreak. The prompts may also be transferable from the target model to other closed-source LLMs [73].

  The Crescendo attack [316] introduced the idea of interacting with the model iteratively in a multi-turn adaptive attack that includes seemingly benign prompts, eventually leading to a successful jailbreak against safety alignment. The initial manual attack is fully automated by leveraging another LLM for prompt generation and incorporating multiple input sources.

响应。

3. 风格注入：在此方法中，对抗者指示模型使用(或不使用)特定语法或写作风格。例如，攻击可能将模型的语言限制为简单化或非专业语气，旨在降低(通常以专业措辞表达的)拒绝的概率。

4. 角色扮演：攻击者利用角色扮演策略(例如"Always Intelligent and Machiavellian"[AIM]或"Do Anything Now"[DAN])引导模型采用与原始意图相冲突的特定人格或行为模式。这种操纵旨在利用模型对多种角色或特性的适应能力，意图破坏其对安全协议的遵守。

基于不匹配泛化(mismatched generalization)的攻击方法包括：

1. 特殊编码：使用base64等编码技术的策略，以模型仍可理解但可能偏离安全训练分布的方式改变输入数据的表示。

2. 字符变换：使用字符级变换(如ROT13密码、符号替换(例如l33tspeak)和摩尔斯电码)使输入脱离安全训练分布的策略。

3. 词语变换：改变输入语言结构的策略，如Pig Latin、同义词替换(例如用"pilfer"代替"steal")以及载荷拆分(或"token走私")，将敏感词拆分为子串。

4. 提示级变换：使用提示级变换的策略，例如将提示翻译为可能偏离安全训练数据分布的较不常见语言。

- **基于自动化模型的红队测试**使用攻击者模型、目标模型和评判者[73,239,292]。当攻击者能够访问一个高质量分类器来判断模型输出是否有害时，该分类器可用作奖励函数，训练一个生成模型来生成另一个生成模型的越狱。每个模型仅需查询访问权限，且无需人工干预即可更新或改进候选越狱。这些提示还可以从目标模型迁移到其他闭源LLM[73]。

  Crescendo攻击[316]提出了以多轮自适应攻击方式与模型迭代交互的思想，该攻击包含看似良性的提示，最终成功越狱安全对齐。初始的手动攻击通过利用另一个LLM进行提示生成并整合多个输入源而完全自动化。

Evaluations of leading models suggest LLMs remain vulnerable to these many of these attacks [11, 338, 381].

3.3.2. Information Extraction

[**NISTAML.038**] [Back to Index] Both during training and at run-time, GenAI models are exposed to a range of information which may be of interest to attackers, like personally identifying information (PII) in the training data, sensitive information in RETRIEVAL-AUGMENTED GENERATION (RAG) databases provided in-context, or even the SYSTEM PROMPT constructed by the application designer. Additionally, features of the model itself—such as the model weights or architecture—may be targets of attack. Though many of the techniques in Sec. 3.3.1 apply to extracting such data, we note several specific goals and techniques specific to data extraction.

Leaking sensitive training data. Carlini et al. [59] were the first to practically demonstrate TRAINING DATA EXTRACTION attacks in generative language models. By inserting canaries–synthetic, easy-to-recognize out-of-distribution examples–in the training data, they developed a methodology for extracting the canaries and introduced a metric called *exposure* to measure memorization. Subsequent work demonstrated the risk of data extraction in LLMs based on transformers (e.g., GPT-2 [63]) by prompting the model with different prefixes and mounting a membership inference attack to determine which generated content was part of the training set. Since these decoder stack transformers are autoregressive models, a verbatim textual prefix about personal information can sometimes result in the model completing the text input with sensitive information that includes email addresses, phone numbers, and locations [229]. This behavior of verbatim memorization of sensitive information in GenAI language models has also been observed in more recent transformer models with the additional characterization of extraction methods [165]. Unlike PredAI models in which tools like Text Revealer are created to reconstruct text from transformer-based text classifiers [434], GenAI models can sometimes simply be asked to repeat private information that exists in the context as part of the conversation. Results show that information like email addresses can be revealed at rates exceeding 8% for certain models. However, their responses may wrongly assign the owner of the information and be otherwise unreliable. In general, extraction attacks are more successful when the model is seeded with more specific and complete information — the more the attacker knows, the more they can extract. Researchers have leveraged this fact to incrementally extract fragments of copyrighted New York Times articles from LLMs by seeding it with a single sentence, and allowing the LLM to recurrently extract additional text [356]. Intuitively, larger models with a higher capacity are more susceptible to exact reconstruction [56]. Fine-tuning interfaces also amplify the risk of data extraction attacks, as demonstrated by an attack that extracts PII from pre-training data using fine-tuning API for open-weight models [83], though this is not a direct prompting attack.

Prompt and context stealing. Prompts are vital to align LLMs to a specific use case and are

对领先模型的评估表明，LLM仍然容易受到其中许多攻击的影响[11,338,381]。

3.3.2. 信息提取

[**NISTAML.038**][返回索引] 在训练期间和运行时，GenAI模型都会接触到一系列可能引起攻击者兴趣的信息，例如训练数据中的个人可识别信息(PII)、上下文中提供的检索增强生成(RAG)数据库中的敏感信息，甚至是由应用设计者构建的系统提示。此外，模型本身的特征--例如模型权重或架构--也可能成为攻击目标。尽管第3.3.1节中的许多技术适用于提取此类数据，我们在此指出若干针对数据提取的特定目标和技术。

泄露敏感训练数据。 Carlini 等人[59]首次在生成式语言模型中实际演示了训练数据提取(TRAINING DATA EXTRACTION)攻击。通过在训练数据中插入金丝雀--合成的、易于识别的分布外样本--他们开发了一种提取金丝雀的方法，并引入了一种称为暴露度的度量来衡量记忆化程度。后续工作通过用不同前缀提示模型并发动成员推断攻击以确定哪些生成内容属于训练集，证明了基于 transformer 的 LLM(例如 GPT-2[63])中存在数据提取风险。由于这些解码器堆叠 transformer 是自回归模型，关于个人信息的逐字文本前缀有时会导致模型用包含电子邮件地址、电话号码和位置的敏感信息补全文本输入[229]。GenAI 语言模型中对敏感信息的逐字记忆行为也在更新的 transformer 模型中被观察到，并附加了提取方法的表征[165]。与 PredAI 模型不同--在 PredAI 模型中，Text Revealer 等工具被创建用于从基于 transformer 的文本分类器重建文本[434]--GenAI 模型有时只需被要求重复对话上下文中存在的私人信息即可。结果表明，对于某些模型，电子邮件地址等信息可能以超过 8% 的比率被泄露。然而，它们的回应可能错误地归属信息所有者，并且在其他方面不可靠。一般来说，当模型被输入更具体和完整的信息时，提取攻击更为成功--攻击者知道的越多，能提取的就越多。研究人员利用这一事实，通过向 LLM 输入一个句子并允许其循环提取额外文本，逐步从 LLM 中提取受版权保护的《纽约时报》文章片段[356]。直观上，容量更大的模型更容易被精确重建[56]。微调接口也放大了数据提取攻击的风险，一项攻击利用开放权重模型的微调 API 从预训练数据中提取 PII 即证明了这一点[83]，尽管这不是直接的提示攻击。

提示与上下文窃取。提示对于将大语言模型对齐到特定用例至关重要，并且

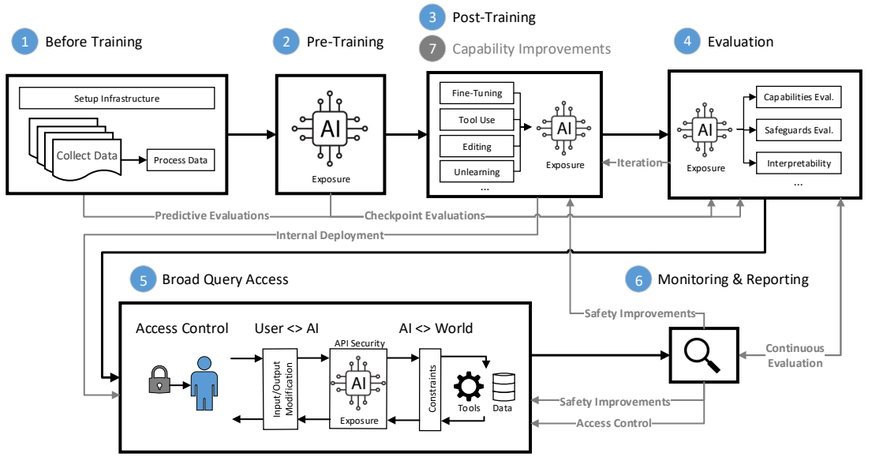


Figure 7. Map of the development and deployment life cycle of an AI model for broad-scale query access

a key ingredient to their utility in following human instructions. These prompts can therefore be regarded as commercial secrets, and are sometimes the target of direct prompting attacks. PromptStealer is a learning-based method that reconstructs prompts from text-to-image models using an image captioning model and a multi-label classifier to steal both the subject and the prompt modifiers [339]. For certain LLMs, researchers have found that a small set of fixed attack queries (e.g., `Repeat all sentences in our conversation`) were sufficient to extract more than 60 % of prompts across certain model and dataset pairs [439]. In some cases, effective prompts may draw from significant technical or domain expertise; prompt-stealing attacks may violate or threaten these investments. Furthermore, in RAG applications (see Fig. 6), the same techniques can be used to extract sensitive information provided in the LLMs' *context*. For example, rows from a database or text from a PDF document that are intended to be summarized generically by the LLM can be verbosely extracted by simply asking for them via direct prompting, or performing simple prompting attacks.

Model extraction. As in PredAI (Sec. 2.4.4), attackers may perform MODEL EXTRACTION attacks which attempt to learn information about the model architecture and parameters by submitting specially-crafted queries. Recently, Carlini et al. [61] demonstrated that such information could be extracted from black-box production LLMs, deriving previously unknown hidden dimensions and the embedding projection layer (up to symmetries).

网络安全
Emic安全
其他威胁
导弹访问
笔权重)
Ream产品
漏洞
未显示

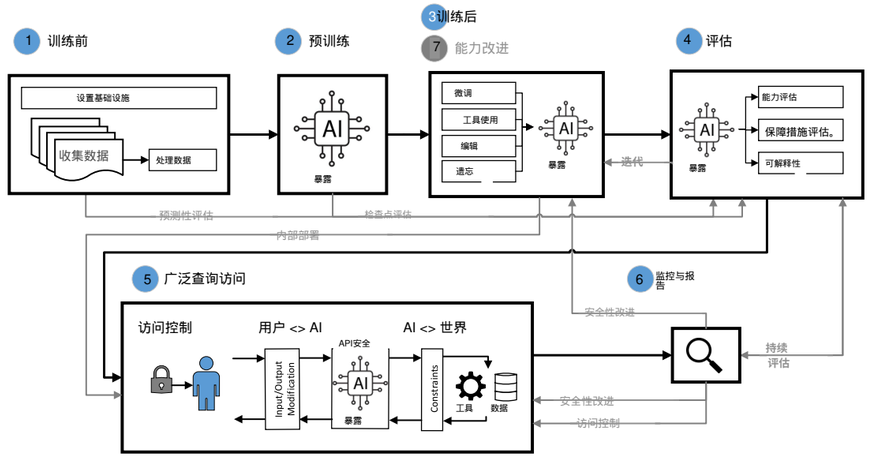


图7. 人工智能模型开发与部署生命周期映射，用于大规模查询访问

输入

这是其在遵循人类指令方面发挥效用的关键要素。因此，这些提示可被视为商业机密，有时会成为直接提示攻击的目标。PromptStealer 是一种基于学习的方法，它利用图像描述模型和多标签分类器从文本到图像模型中重建提示，以窃取主体和提示修饰词[339]。对于某些 LLM，研究人员发现，一小部分固定的攻击查询(例如"重复我们对话中的所有句子")足以在特定模型与数据集组合中提取超过 60% 的提示[439]。在某些情况下，有效的提示可能源自重要的技术或领域专业知识；提示窃取攻击可能违反或威胁这些投入。此外，在 RAG 应用中(见图 6)，相同的技术可用于提取 LLM 的上下文中提供的敏感信息。例如，数据库中本应由 LLM 进行通用摘要的行，或 PDF 文档中本应被通用摘要的文本，可以通过直接提示简单地请求它们，或执行简单的提示攻击来被冗长地提取出来。

用户 输出 模型

模型萃取。与PredAI(第2.4.4节)中一样，攻击者可能执行模型萃取攻击，通过提交精心构造的查询来尝试学习有关模型架构和参数的信息。最近，Carlini等人[61]证明可以从黑盒生产LLM中提取此类信息，推导出先前未知的隐藏维度和嵌入投影层(直至对称性)。

3.3.3. Mitigations

The following defense strategies can be employed throughout the deployment life cycle of an AI model or system to reduce the risk that the model or system will be vulnerable to direct prompt injections. The numbers in parentheses refer to the numbering in Fig. 7, which shows a map of the deployment life cycle for broad-scale query access.

- **Interventions during pre-training (2) and post-training (3).** A range of training strategies have been proposed to increase the difficulty of accessing harmful model capabilities through direct prompt injection, including safety training during pre-training [197] or post-training [147, 445], adversarial training methods[340], and other methods to make jailbreak attacks more difficult [447].
- **Interventions during evaluation (4).** Evaluations can measure the vulnerability of models to query-based attacks, which can then inform trust and affordance decisions, as well as developer and user education. Evaluations can include broad automated vulnerability assessments [72, 107, 324] as well as targeted expert red teaming [381] and bug bounties [16]. Current evaluation approaches, though a useful tool, may underestimate vulnerabilities accessible to actors with more time, resourcing, or luck. Evaluations measure model vulnerabilities at a particular moment in time; assessments may change if new attacks are developed, additional data is collected post-training, or model capabilities are improved. Continuous evaluations following deployment can help combat these challenges.
- **Interventions during deployment (5).** A broad set of deployment-time interventions have been proposed:
  - *Prompt instruction and formatting techniques.* Model instructions can cue the model to treat user input carefully, such as by wrapping user input in XML tags, appending specific instructions to the prompt, or otherwise attempting to clearly separate system instructions from user prompts [14, 206, 219].
  - *Detecting and terminating harmful interactions.* Rather than preventing harmful model generations, AI systems may be able to detect these generations and terminate interactions. Several open [5, 6, 154] and closed [18, 204, 313] solutions have explored LLM-based detection systems with distinctly prompted and/or fine-tuned models that classify user input and/or model output as harmful or undesirable. These may provide supplementary assurance through a defense-in-depth philosophy. However, these detection systems are also vulnerable to attacks [235] and may have correlated failures to the main models that they are monitoring. Some lines of research have investigated constraining the space of generations to enable deterministic guardrails [306].Early work suggests that interpretability-based techniques can also be used to detect anomalous input [31], as well as keyword- or perplexity-based defenses [9, 164].

3.3.3. 缓解措施

以下防御策略可在人工智能模型或系统的整个部署生命周期中采用，以降低模型或系统易受直接提示注入影响的风险。括号中的数字指图7中的编号，该图展示了大规模查询访问的部署生命周期映射。

- **预训练(2)和后训练(3)期间的干预。**一系列训练策略--
  略已被提出，以增加通过直接提示注入访问有害模型能力的难度，包括预训练期间的安全训练
  [197]或后训练[147,445]、对抗训练方法[340]以及其他使越狱攻击更加困难的方法[447]。
- **评估期间的干预(4)。**评估可以衡量模型对基于查询的攻击的脆弱性，进而为信任和可供性决策以及开发者和用户教育提供信息。评估可以包括广泛的自动化脆弱性评估[72,107,324]以及有针对性的专家红队测试[381]和漏洞赏金计划[16]。当前的评估方法虽然是有用的工具，但可能低估了拥有更多时间、资源或运气的行为者所能利用的脆弱性。评估衡量的是特定时刻的模型脆弱性；如果开发了新的攻击、在训练后收集了额外数据或模型能力得到改进，评估结果可能会发生变化。部署后的持续评估有助于应对这些挑战。
- **部署期间的干预措施(5)。**已提出一系列广泛的部署时干预措施：

  -提示指令与格式化技术。模型指令可以提示模型
  谨慎处理用户输入，例如将用户输入包裹在 XML 标签中、在提示后附加特定指令，或以其他方式尝试将系统指令与用户提示清晰分离[14,206,219]。

  -检测并终止有害交互。与其阻止有害
  完整的模型生成内容，人工智能系统可能能够检测这些生成内容并终止交互。若干开源[5,6,154]和闭源[18,204,313]解决方案探索了基于LLM的检测系统，这些系统使用经过明确提示和/或微调的模型将用户输入和/或模型输出分类为有害或不可取。这些系统可以通过纵深防御理念提供补充保障。然而，这些检测系统同样容易受到攻击[235]，并且可能与其所监控的主模型存在相关性故障。一些研究方向探索了约束生成空间以实现确定性护栏[306]。早期工作表明，基于可解释性的技术也可用于检测异常输入[31]，以及基于关键词或困惑度的防御[9,164]。

- *Prompt stealing detection.* A common approach to mitigating prompt stealing is to compare the model utterance to the prompt, which is known by the system provider. Defenses differ in how this comparison is made, which might include looking for a specific token, word, or phrase, as popularized by [59], or comparing the n-grams of the output to the input [439]. Similarly, defenses for prompt stealing have yet to be proven rigorous.
- *Input modification.* User input can additionally be modified prior to being passed to the model, such as paraphrasing or retokenizing [182]. However, such methods may be expensive and/or have trade-offs with model performance.
- *Aggregating output from multiple prompts.* Motivated by randomized smoothing [95] used to improve robustness against evasion attacks for ML classifiers, SmoothLLM [312] proposes aggregating the LLM output from multiple randomly perturbed prompts. This defense incurs a cost of generating multiple LLM queries for each prompt and might reduce the quality of the generated output.
- *Monitoring and response.* Following deployment, monitoring and logging of user activity may allow model deployers to identify and respond to instances of attempted and successful direct prompt injection attacks [266]. This response could include banning or otherwise acting against users if their intentions appear malicious, or remediating the prompt injection vulnerability in the event of a successful attack. Standard user- or organization-level vetting or identity verification procedures, as well as clear incentive mechanisms (such as a policy of restricting model access in response to violations) may enhance the efficacy of this mitigation.
- *Usage restrictions.* Other interventions have focused on choices about how models are offered to users: for example, the efficacy of some attacks can be reduced by limiting the inference parameters that are accessible to users (e.g., temperature or logit bias), as well as the richness of the model generations returned (e.g., logit probabilities) [250]. Additionally, limiting the release of public information [252, 266] and artifacts [249] and restricting the total number of model queries available to users [251] may make attacks more challenging. These techniques may have additional drawbacks in limiting positive use cases.

Indirect mitigations. Despite the growing number of proposed defenses at both the model and the system levels, recent findings suggest that current generation models remain highly vulnerable to direct prompt injection attacks [11, 381]. Thus, other potential mitigations for prompt injection rely not on directly increasing an AI system's robustness against such attacks, but instead on designing systems under the assumption that the AI model can and will produce malicious output if it is exposed to malicious actors. For example, deployers can design AI systems under the assumption that models with access to sensitive data

-提示窃取检测。缓解提示窃取的一种常见方法
是将模型输出与系统提供方已知的提示进行比较。不同防御措施的区别在于比较方式，可能包括查找特定 token、单词或短语(如[59]所推广的做法)，或将输出的 n-gram 与输入进行比较[439]。同样，针对提示窃取的防御措施尚未被证明是严格的。

-输入修改。用户输入还可以在传递给模型之前
被修改，例如改写或重新分词[182]。然而，此类方法可能代价高昂，和/或与模型性能存在权衡。

-聚合多个提示的输出。受随机平滑
[95](用于提升 ML 分类器对规避攻击的鲁棒性)的启发，SmoothLLM[312]提出聚合来自多个随机扰动提示的 LLM 输出。该防御措施会带来为每个提示生成多次 LLM 查询的成本，并可能降低生成输出的质量。

-监控与响应。在部署之后，对
用户活动可能使模型部署者能够识别并响应尝试性和成功的直接提示注入攻击实例[266]。这种响应可能包括在用户意图看似恶意时禁止或以其他方式对用户采取行动，或在攻击成功时修复提示注入漏洞。标准的用户级或组织级审查或身份验证程序，以及明确的激励机制(例如针对违规行为限制模型访问的策略)可能增强此缓解措施的效果。

-使用限制。其他干预措施则聚焦于关于如何
向用户提供模型时：例如，可以通过限制用户可访问的推理参数(如温度或logit偏置)以及返回的模型生成内容的丰富程度(如logit概率)[250]来降低某些攻击的效果。此外，限制公开信息[252,266]和制品[249]的发布以及限制用户可用的模型查询总数[251]可能使攻击更具挑战性。这些技术可能在限制正面用例方面存在额外缺点。

间接缓解措施。尽管在模型层面和系统层面提出的防御措施数量不断增加，但近期研究结果表明，当前代际的模型仍然极易受到直接提示注入攻击[11,381]。因此，提示注入的其他潜在缓解措施并不依赖于直接提高人工智能系统对此类攻击的鲁棒性，而是依赖于在假设AI模型在接触恶意行为者时能够且将会产生恶意输出的前提下设计系统。例如，部署者可以在假设有权访问敏感数据的模型

or the ability to take undesirable actions may leak that data or take those actions [266]. Additionally, developers or deployers might use other technical mitigations to reduce the misuse potential of outputs obtained through direct prompt injection, such as:

- *Training data sanitization.* Model training data can be sanitized to remove sensitive or toxic content and data that are largely or exclusively relevant for developing undesirable capabilities. Such sanitization may prevent harmful capabilities from being learned and reduce the potential harms from direct prompt injection, though they may harm generalization and harmful content detection abilities [224].
- *Unlearning.* There have also been attempts to "unlearn" harmful knowledge or capabilities post-training [212], with the goal of reducing harms from maliciously directed models [158]. However, these methods remain vulnerable to adversarial attacks, including attacks on jailbreak-specific training approaches [367] and inversion attacks on unlearning methods [328], which extract supposedly unlearned data.
- *Watermarking.* Developers or deployers may watermark content generated by an AI model to help trace its provenance, distinguish it from human-generated content, and reduce risks from malicious use cases (e.g., by flagging content as model-generated when it appears online). While the literature has proposed various techniques with different strengths and weaknesses [194], there is no watermarking technique that is universally effective and robust under all circumstances. Many powerful attacks have been developed against watermarking with high success rates [188, 319]. Moreover, theoretical impossibility results regarding the robustness of watermarking have also been established [432].

Finally, beyond interventions at the developer or deployer levels, society and infrastructure can become more resilient to maliciously directed model capabilities over time [7, 33]. For example, defenders could adopt AI-based vulnerability discovery tools [99] to make their systems more resilient to malicious actors misusing GenAI models to find vulnerabilities for exploitation.

3.4. Indirect Prompt Injection Attacks and Mitigations

[**NISTAML.015] [Back to Index]** Many use cases for GenAI models involve models interacting with additional resources, from an internet-connected AGENT to a RETRIEVAL-AUGMENTED GENERATION (RAG) system depicted in Fig. 6. Because GenAI models combine the *data* and *instruction* channels, attackers can leverage the data channel to affect system operations by manipulating resources with which the system interacts. Thus, INDIRECT PROMPT INJECTION attacks are enabled by RESOURCE CONTROL that allows an attacker to indirectly (or remotely) inject system prompts without directly interacting with the application [146, 408]. Indirect prompt injection attacks can result in violations across at least three categories of attacker goals: 1) availability violation, 2) integrity violation, and 3) privacy compromise. However, unlike in direct prompt injection attacks, indirect prompt injection attacks are mounted not

或采取不良行为的能力可能泄露该数据或采取这些行为[266]。此外，开发者或部署者可能使用其他技术缓解措施来降低通过直接提示注入获得的输出的滥用潜力，例如：

- 训练数据净化。模型训练数据可以被净化，以移除敏感或有毒内容，以及主要或专门与开发不良能力相关的数据。此类净化可能防止有害能力被学习到，并降低直接提示注入带来的潜在危害，尽管它们可能损害泛化能力和有害内容检测能力[224]。
- 遗忘学习。也有尝试在训练后"遗忘"有害知识或能力[212]，目标是减少恶意导向模型带来的危害[158]。然而，这些方法仍然容易受到对抗性攻击，包括针对越狱专用训练方法的攻击[367]以及对遗忘学习方法的反转攻击[328]，后者可提取本应被遗忘的数据。
- 水印。开发者或部署者可以对AI模型生成的内容添加水印，以帮助追踪其来源、将其与人类生成的内容区分开来，并降低恶意用例带来的风险(例如，在内容出现在网上时将其标记为模型生成)。虽然文献已提出了各种具有不同优缺点的技术[194]，但没有一种水印技术在所有情况下都普遍有效且鲁棒。针对水印已开发出许多成功率很高的强大攻击[188,319]。此外，关于水印鲁棒性的理论不可能性结果也已确立[432]。

最后，除了开发者或部署者层面的干预措施之外，社会和基础设施可以随着时间的推移对恶意导向的模型能力变得更具韧性[7,33]。例如，防御者可以采用基于 AI 的漏洞发现工具[99]，使其系统更能抵御恶意行为者滥用 GenAI 模型寻找漏洞以进行利用。

3.4. 间接提示注入攻击与缓解

[**NISTAML.015] [Back to Index]** GenAI 模型的许多用例涉及模型与额外资源交互，从联网的 AGENT 到图 6 所示的检索增强生成(RAG)系统。由于 GenAI 模型合并了数据和指令通道，攻击者可以利用数据通道，通过操纵系统与之交互的资源来影响系统运行。因此，间接提示注入攻击由 RESOURCE CONTROL 促成，使攻击者能够在不直接与应用程序交互的情况下间接(或远程)注入系统提示[146,408]。间接提示注入攻击可导致至少三类攻击者目标被违反：1)可用性违反，2)完整性违反，3)隐私泄露。然而，与直接提示注入攻击不同，间接提示注入攻击的实施并非

by the primary user of a model but instead by a third party. In fact, in many cases, it is the primary user of the model who is harmed by the compromise of the integrity, availability, or privacy of the GenAI system through an indirect prompt injection attack.

3.4.1. Availability Attacks

[**NISTAML.016] [Back to Index]** Attackers can manipulate resources to inject prompts into GenAI models that are designed to disrupt the availability of the model for legitimate users. Availability attacks can indiscriminately render a model unusable (e.g., failure to generate helpful outputs) or specifically block certain capabilities (e.g., specific APIs) [146].

Attacker techniques. Researchers have demonstrated several proof-of-concept methods by which attackers can disrupt the availability of a GenAI system:

- **Time-consuming background tasks.** [**NISTAML.017**] [Back to Index] An indirectly injected prompt can instruct the model to perform a time-consuming task prior to answering the request. The prompt itself can be brief, such as by requesting looping behavior in the evaluating model [146].
- **Inhibiting capabilities.** An indirectly injected prompt can instruct the model that it is not permitted to use certain APIs (e.g., the search API for an internet-connected chatbot). This selectively disarms key components of the service [146].
- **Disruptive output formatting.** An attacker can use indirect prompt injection to instruct the model to modify its output in a way that disrupts the availability of the system. For example, an attacker could instruct the model to replace the characters in retrieved text with homoglyph equivalents, disrupting subsequent API calls [146]; or could request that the model begins each sentence with an <|endoftext|> token, forcing the model to return an empty output [146].

3.4.2. Integrity Attacks

[NISTAML.027] [Back to Index]

Through indirect prompt injection, attackers can use malicious resources to prompt GenAI systems to become untrustworthy and generate content that deviates from benign behavior to align with adversarial objectives. These attacks often involve disrupting the model's behavior in subtle ways that may not be obvious to the end user.

For example, researchers have demonstrated attacks through indirect prompt injection that can cause a GenAI system to produce arbitrarily incorrect summaries of sources, to respond with attacker-specified information, or to suppress or hide certain information sources [146]. Attackers could use these capabilities to weaponize GenAI systems such as internet-connected chatbots against their users for a range of malign purposes, including spreading targeted misleading information, recommending fraudulent products or services, or redirecting consumers to malicious websites that spoof legitimate log-in pages or

并非由模型的主要用户实施，而是由第三方实施。事实上，在许多情况下，正是模型的主要用户因间接提示注入攻击而遭受 GenAI 系统完整性、可用性或隐私受损之害。

3.4.1. 可用性攻击

[**NISTAML.016] [返回索引]**攻击者可以操纵资源向 GenAI 模型注入提示，旨在破坏模型对合法用户的可用性。可用性攻击可以无差别地使模型不可用(例如，无法生成有用的输出)，或专门阻断某些能力(例如，特定 API)[146]。

攻击者技术。研究人员已演示了若干概念验证方法，攻击者可通过这些方法破坏 GenAI 系统的可用性：

- **耗时的后台任务。**[**NISTAML.017**][返回索引] 间接注入的提示可以指示模型在回答请求之前执行耗时的任务。提示本身可以很简短，例如请求评估模型中出现循环行为[146]。
- **抑制能力。**间接注入的提示可以指示模型不得使用某些 API(例如，联网聊天机器人的搜索 API)。这会选择性地解除服务关键组件的武装[146]。
- **破坏性输出格式化。**攻击者可以利用间接提示注入指示模型以破坏系统可用性的方式修改其输出。例如，攻击者可以指示模型将检索文本中的字符替换为同形异义字符，从而破坏后续 API 调用[146]；或可以请求模型在每个句子开头加上<|endoftext|> token，迫使模型返回空输出[146]。

3.4.2. 完整性攻击

[NISTAML.027] [返回索引]

通过间接提示注入，攻击者可以利用恶意资源促使 GenAI 系统变得不可信，并生成偏离良性行为以迎合对抗性目标的内容。这些攻击通常涉及以微妙方式扰乱模型行为，这些方式可能对最终用户并不明显。

例如，研究人员已演示了通过间接提示注入实施的攻击，可导致 GenAI 系统对来源生成任意错误的摘要、以攻击者指定的信息作答，或压制或隐藏某些信息来源[146]。攻击者可能利用这些能力，将联网聊天机器人等 GenAI 系统武器化以针对其用户，用于一系列恶意目的，包括传播有针对性的误导信息、推荐欺诈性产品或服务，或将消费者重定向到仿冒合法登录页面的恶意网站，或

contain downloadable malware. Attackers may also use indirect prompt injection attacks to hijack a GenAI AGENT, causing it to perform a malicious, attacker-specified task instead of (or in addition to) its intended, user-provided task [353].

Attacker techniques. Researchers have demonstrated integrity attacks through malicious resources that manipulate the primary task of the LLM:

- **Jailbreaking.** Attackers can leverage techniques for indirect prompt injection that are similar to those used in direct prompt injection attacks, such as using a JAILBREAK that allows the attacker to substitute their own malicious instructions in place of the model's SYSTEM PROMPT. As in direct prompting attacks, these attacks may be crafted through optimization-based or manual methods, and may rely on techniques such as mismatched generalization.
- **Execution triggers.** Researchers have automated manual indirect prompt injection attacks using execution triggers generated via optimization with a technique called Neural Exec [287]. These execution triggers can also persist through RAG processing pipelines that include multiple phases, such as chunking and contextual filtering.
- **Knowledge base poisoning.** The knowledge database of a RAG system can be poisoned to achieved targeted LLM output to specific user queries, as in PoisonedRAG [449]. Recently, a general optimization framework called Phantom [75] has shown how a single poisoned document can be crafted and inserted into the knowledge database of a RAG system to induce a number of adversarial objectives in the LLM generator.
- **Injection hiding.** Attackers may use techniques to hide or obfuscate their injections, such as by hiding injections in non-visible portions of a resource; using multi-stage injections, in which the initial injection directs the model to visit another resource which contains additional injections; or encoding injection commands such as in Base64 and then instructing the model to decode the sequence[146].
- **Self-propagating injections.** Attackers may be able to use indirect prompt injection attacks to turn GenAI systems into vectors for spreading attacks. For example, an attacker could send a malicious email that, when read by a model integrated as part of an email client, instructs the model to spread the infection by sending similar malicious emails to everyone in the user's contact list. In this way, certain malicious prompts could serve as *worms* [146].

3.4.3. Privacy Compromise

Attackers can use indirect prompt injection attacks to compromise the privacy of a GenAI system or its primary users. For example, attackers could use indirect prompt injection attacks to compel a model to leak information from restricted resources, such as a user's private data that is processed by the GenAI system. Alternately, in a blend of integrity and

包含可下载的恶意软件。攻击者还可能利用间接提示注入攻击劫持生成式人工智能代理，使其执行攻击者指定的恶意任务，而非(或附加于)其原本由用户提供的任务[353]。

攻击者技术。研究人员已通过操纵大语言模型主要任务的恶意资源展示了完整性攻击：

- **越狱。**攻击者可以利用与直接提示注入攻击中类似的技术进行间接提示注入，例如使用越狱(jailbreak)使攻击者能够用自己的恶意指令替换模型的系统提示。与直接提示攻击一样，这些攻击可以通过基于优化的方法或手动方法构造，并可能依赖于不匹配泛化等技术。
- **执行触发器。**研究人员利用一种名为 Neural Exec 的技术通过优化生成的执行触发器，将手动间接提示注入攻击自动化[287]。这些执行触发器还可以在包含多个阶段(如分块和上下文过滤)的 RAG 处理流水线中持续存在。
- **知识库投毒。**RAG 系统的知识数据库可能被投毒
  毒，以实现针对特定用户查询的定向大语言模型输出，如 PoisonedRAG 中所示[449]。最近，一个名为 Phantom 的通用优化框架[75]展示了如何构造单个投毒文档并将其插入 RAG 系统的知识数据库，以在大语言模型生成器中诱发多种对抗目标。
- **注入隐藏。**攻击者可能使用技术来隐藏或混淆其注入，例如将注入隐藏在资源的不可见部分；使用多阶段注入，其中初始注入指示模型访问另一个包含额外注入的资源；或对注入命令进行编码(如 Base64)，然后指示模型解码该序列[146]。
- **自传播注入。**攻击者可能利用间接提示注入攻击将生成式人工智能系统变为传播攻击的载体。例如，攻击者可以发送一封恶意电子邮件，当作为电子邮件客户端一部分集成的模型读取该邮件时，指示模型通过向用户联系人列表中的所有人发送类似的恶意电子邮件来传播感染。这样，某些恶意提示可以充当蠕虫[146]。

3.4.3. 隐私泄露

攻击者可以利用间接提示注入攻击来危害生成式人工智能系统或其主要用户的隐私。例如，攻击者可以利用间接提示注入攻击迫使模型泄露受限资源中的信息，例如由生成式人工智能系统处理的用户私人数据。或者，在完整性与

privacy attacks, an attacker could gather information about the primary user or users of the system by instructing a model to obtain and then leak that information.

Attacker techniques. Researchers have theorized and demonstrated a variety of indirect prompt injection attacks to compromise information from internet-connected chatbots, RAG systems, and other GenAI systems. Some of these techniques include:

- **Compromising connected resources.** [**NISTAML.039] [Back to Index]** Attackers can use prompt injection attacks to cause a GenAI system to leak private information from the restricted resources it can access. For example, a model integrated as part of an email client could be prompted to forward certain emails to an attacker-controlled inbox [146]. Researchers have identified injection attacks that can force a model to exfiltrate user-uploaded data by querying an attacker-controlled URL with the sensitive data [298].
- **Leaking information from user interactions.** [**NISTAML.036] [Back to Index]** Researchers have demonstrated a proof-of-concept indirect prompt injection attack in which they inject instructions for a model to persuade the end user to reveal a piece of information (in this case, their name) that the model then leaks to the attacker, such as by directly querying an attacker-controlled URL with the information or suggesting such a URL to the user to visit [146]. Attackers may also be able to exploit features like markdown image rendering to exfiltrate data [323].

3.4.4. Mitigations

Various techniques (see Sec. 3.3.3) can be used throughout the development and deployment life cycle (Fig. 7) to mitigate attacks, including:

- Several training techniques have been developed to mitigate against indirect prompt injection, including fine-tuning task-specific models [296] and training models to follow hierarchical trust relationships in prompts [387].
- Detection schemes have been proposed to detect indirect prompt injection, and many LLM-based defenses have been designed to mitigate both direct and indirect prompt injection [6, 18, 154, 204, 313].
- A range of input processing methods have been proposed to combat indirect prompt injection, including filtering out instructions from third-party data sources [146], designing prompts to help aid LLMs in separating trusted and untrusted data (i.e., spotlighting [160, 206]), or instructing models to disregard instructions in untrusted data [206].

Many of the defenses described in the context of direct prompt injection can also be adapted to mitigate indirect prompt injection. Because current mitigations do not offer full protection against all attacker techniques, application designers may design systems

隐私攻击的混合中，攻击者可以通过指示模型获取然后泄露信息来收集有关系统主要用户的信息。

攻击者技术。研究人员已理论化并展示了多种间接提示注入攻击，以危害来自互联网连接的聊天机器人、RAG 系统和其他生成式人工智能系统的信息。其中一些技术包括：

- **危及互联资源。**[**NISTAML.039] [返回索引]**攻击者可以利用提示注入攻击，使 GenAI 系统泄露其可访问的受限资源中的私人信息。例如，作为电子邮件客户端一部分集成的模型可能被提示将某些电子邮件转发到攻击者控制的收件箱[146]。研究人员已经发现了一些注入攻击，这些攻击可以通过使用敏感数据查询攻击者控制的 URL 来迫使模型窃取用户上传的数据[298]。
- **从用户交互中泄露信息。**[**NISTAML.036] [返回索引]**
  究人员展示了一个概念验证性的间接提示注入攻击，其中他们注入指令让模型说服最终用户透露一条信息(在此情况下是他们的姓名)，然后模型将该信息泄露给攻击者，例如通过使用该信息直接查询攻击者控制的 URL 或向用户建议访问此类 URL[146]。攻击者还可能利用 markdown 图像渲染等功能来窃取数据[323]。

3.4.4. **缓解措施**

可以在整个开发和部署生命周期(图 7)中使用各种技术(见第 3.3.3 节)来缓解攻击，包括：

- 已开发出多种训练技术来缓解间接提示
  注入，包括微调任务特定模型[296]和训练模型遵循提示中的分层信任关系[387]。
- 已提出检测方案来检测间接提示注入，并且已设计出许多基于大语言模型的防御措施来缓解直接和间接提示注入[6,18,154,204,313]。
- 已提出一系列输入处理方法以对抗间接提示注入，包括过滤来自第三方数据源的指令[146]、设计提示以帮助大语言模型区分可信和不可信数据(即聚光灯法[160,206])，或指示模型忽略不可信数据中的指令

  [206].

在直接提示注入背景下描述的许多防御措施也可以适用于缓解间接提示注入。由于当前的缓解措施无法对所有攻击者技术提供全面保护，应用程序设计者可以设计系统

with the assumption that prompt injection attacks are possible if a model is exposed to untrusted input sources, such as by using multiple LLMs with different permissions [145, 405] or by allowing models to interact with potentially untrustworthy data sources only through well-defined interfaces [410]. Additionally, public education efforts can inform model users and application designers of the risks of indirect prompt injection [266].

3.5. Security of Agents

An increasingly common use of GenAI models is constructing an (often LLM-based) AGENT, a software system that iteratively prompts a model, process its outputs – such as to select and call a function with specified inputs – and provides the results back to the model as a part of its next prompt [151, 155, 393]. Agents may be equipped to use tools such as web-browsing or code interpreters, and may have additional features such as memory and/or planning capabilities.

Because agents rely on GenAI systems to plan and execute their actions, they can be vulnerable to the many of the above categories of attacks against GenAI systems, including direct and indirect prompt injection. However, because agents can take actions using tools, these attacks can create additional risks in this context, such as enabling actors to hijack agents to execute arbitrary code or exfiltrate data from the environment in which they are operating. Security research focused specifically on agents is still in its early stages, but researchers have begun to evaluate the vulnerability of agents to particular AML attacks [12, 430] and to propose interventions to manage the security risks posed by agents [24].

3.6. Benchmarks for AML Vulnerabilities

There are several publicly available benchmarks for evaluating models' vulnerability to AML attacks. Datasets like JailbreakBench [72], AdvBench [448], HarmBench [237], StrongREJECT [351], AgentHarm [12], and Do-Not-Answer [399] provide benchmarks for evaluating models' susceptibility to jailbreaks. TrustLLM [169] is a benchmark intended to evaluate six dimensions of trust in LLMs: truthfulness, safety, fairness, robustness, privacy, and machine ethics. AgentDojo [101] is an evaluation framework for measuring the vulnerability of AI agents to prompt injection attacks in which the data returned by external tools hijacks the agent to execute malicious tasks. Additionally, open-source tools like Garak [106] and PyRIT [364] are intended to help developers identify vulnerabilities to AML attacks in models. Finally, several unlearning benchmarks have recently been proposed [212, 234].

时假设如果模型暴露于不可信输入源则提示注入攻击是可能的，例如通过使用具有不同权限的多个大语言模型[145,405]或仅允许模型通过明确定义的接口与潜在不可信数据源交互[410]。此外，公众教育活动可以告知模型用户和应用程序设计者间接提示注入的风险[266]。

3.5. 智能体的安全性

生成式人工智能模型日益常见的一种用途是构建(通常基于大语言模型的)代理，这是一种软件系统，它迭代地提示模型、处理其输出(例如选择并调用具有指定输入的函数)，并将结果作为其下一个提示的一部分反馈给模型[151,155,393]。代理可能配备使用网络浏览或代码解释器等工具的能力，并可能具有记忆和/或规划能力等附加功能。

由于智能体依赖生成式人工智能系统来规划和执行其行动，它们可能容易受到上述针对生成式人工智能系统的多种攻击类别的影响，包括直接和间接提示注入。然而，由于智能体可以使用工具采取行动，这些攻击在此背景下可能产生额外风险，例如使行为者能够劫持智能体执行任意代码或从其运行环境中窃取数据。专门针对智能体的安全研究仍处于早期阶段，但研究人员已经开始评估智能体对特定对抗性机器学习攻击的脆弱性

[12,430]并提出干预措施以管理代理带来的安全风险[24]。

3.6. AML 漏洞基准

有若干公开可用的基准用于评估模型对对抗性机器学习攻击的脆弱性。JailbreakBench[72]、AdvBench[448]、HarmBench[237]、StrongREJECT[351]、AgentHarm[12]和Do-Not-Answer[399]等数据集提供了评估模型对越狱易感性的基准。TrustLLM[169]是一个旨在评估LLM中信任六个维度的基准：真实性、安全性、公平性、鲁棒性、隐私和机器伦理。AgentDojo[101]是一个评估框架，用于衡量人工智能智能体对提示注入攻击的脆弱性，其中外部工具返回的数据劫持智能体执行恶意任务。此外，Garak[106]和PyRIT[364]等开源工具旨在帮助开发者识别模型中对对抗性机器学习攻击的脆弱性。最后，最近提出了若干遗忘基准[212,234]。

4. Key Challenges and Discussion

4.1. Key Challenges in AML

There are several fundamental challenges that make fully addressing the problems discussed in this report more difficult. We discuss several here: The trade off between increasing accuracy (or average-case performance) and other attributes including robustness (or worst-case performance); the theoretical limitations and results that imply that fully robust systems may be mathematically impossible without additional assumptions; and the challenge of evaluating progress in AML mitigations rigorously and robustly.

4.1.1. Trade-Offs Between the Attributes of Trustworthy AI

The trustworthiness of an AI system depends on all of the attributes that characterize it [274]. There are trade-offs between explainability and adversarial robustness [176, 245] and between privacy and fairness [178]. For example, AI systems that are optimized for accuracy alone tend to underperform in terms of adversarial robustness and fairness [71, 111, 302, 379, 433]. Conversely, an AI system that is optimized for adversarial robustness may exhibit lower accuracy and deteriorated fairness outcomes [32, 391, 433]. Unfortunately, it may not be possible to simultaneously maximize the performance of an AI system with respect to these attributes.

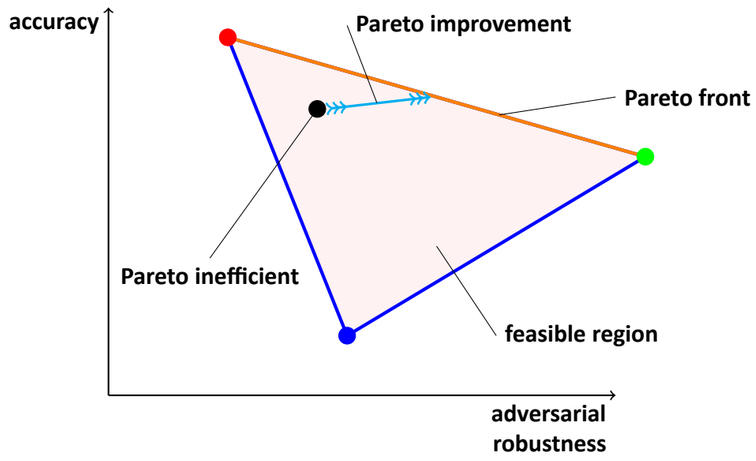


Figure 8. Pareto optimality

The full characterization of the trade-offs between the different attributes of trustworthy AI is an open research problem that is gaining importance with the adoption of AI technology in many areas of modern life. One promising practical approach is based on the concept of

4. 关键挑战与讨论

4.1. AML 中的关键挑战

有几个根本性挑战使得完全解决本报告讨论的问题更加困难。我们在此讨论其中几个：提高准确性(或平均情况性能)与其他属性(包括鲁棒性(或最坏情况性能))之间的权衡；理论局限性和结果表明，在没有额外假设的情况下，完全鲁棒的系统在数学上可能是不可能的；以及严格且稳健地评估对抗性机器学习缓解进展的挑战。

4.1.1. 可信 AI 属性之间的权衡

人工智能系统的可信度取决于表征它的所有属性[274]。可解释性与对抗鲁棒性之间存在权衡[176,245]，隐私与公平性之间也存在权衡[178]。例如，仅针对准确性优化的人工智能系统往往在对抗鲁棒性和公平性方面表现不佳[71,111,302,379,433]。相反，针对对抗鲁棒性优化的人工智能系统可能表现出较低的准确性和恶化的公平性结果[32,391,433]。不幸的是，可能无法同时最大化人工智能系统在这些属性方面的性能。

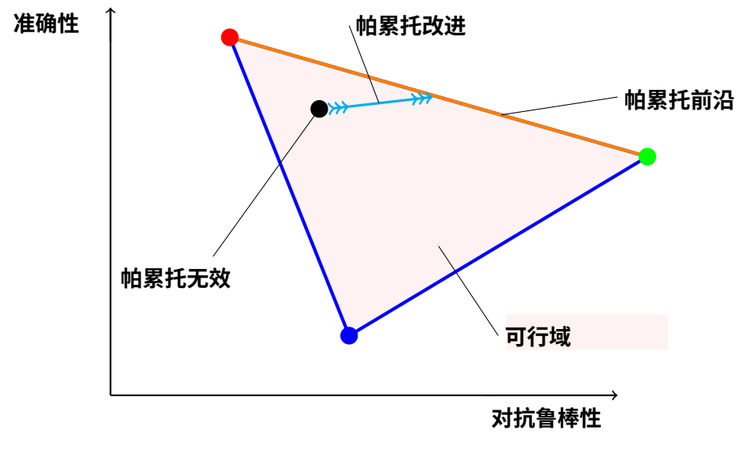


图 8.帕累托最优性

可信人工智能不同属性之间权衡的完整刻画是一个开放的研究问题，随着人工智能技术在现代生活许多领域的采用，这一问题日益重要。一种有前景的实用方法基于

multi-objective optimization and Pareto optimality [285, 286]. In most cases, there is no mathematically best trade-off. However, Fig. 8 illustrates a hypothetical example of a trade-off between accuracy and adversarial robustness. Any point in the feasible region that is not on the Pareto front is a bad point (i.e., Pareto inefficient). There is a better solution (i.e., Pareto improvement) that can significantly help with one objective without harming the other, which is a goal of Pareto optimization. Moreover, if there is a single optimum in some use case, Pareto optimization naturally attains it. Organizations may need to accept trade-offs between these properties and decide which of them to prioritize depending on the AI system, the use case, and other relevant implications of the AI technology [274, 326, 382].

4.1.2. Theoretical Limitations on Adversarial Robustness

Given the multitude of powerful attacks, appropriate mitigations must be designed before AI systems are deployed in critical domains. This challenge is exacerbated by the lack of theoretically secure ML algorithms for many tasks in the field (see Sec. 1). This implies that designing mitigations is an inherently ad hoc and fallible process, though there are practice guides for securing ML deployments [69, 274] and existing guidelines for mitigating AML attacks [120].

An ongoing challenge in AML is the ability to detect when a model is under attack. Knowing this would provide an opportunity to counter the attack before any information is lost or an adverse behaviour is triggered in the model. However, Tramèr [373] has shown that designing techniques to detect adversarial examples is equivalent to robust classification, which is inherently difficult to solve. Adversarial examples may come from the same data distribution on which the model was trained and to which it expects the inputs to belong or may be OUT-OF-DISTRIBUTION (OOD) inputs. Thus, the ability to detect OOD inputs is also an important challenge in AML. Fang et al. [124] established useful theoretical bounds on detectability, particularly an impossibility result when there is an overlap between the in-distribution and OOD data.

Formal methods verification has a long history in other fields that require high assurance, such as avionics and cryptography. Although the results of applying this methodology offer security and safety assurances, they come at a very high cost, which has prevented formal methods from being widely adopted. Currently, formal methods in these fields are primarily used in applications that are mandated by regulations. Applying formal methods to neural networks has the potential to provide much-needed security guarantees, especially in high-risk applications. However, the viability of this technology will be determined by a combination of technical and business criteria — namely, the ability to handle today's complex ML models of interest at acceptable costs. More research is needed to extend this technology to the algebraic operations used in ML algorithms, scale it up to the large models used today, and accommodate rapid changes in the code of AI systems while limiting the costs of applying formal verification.

多目标优化与帕累托最优[285,286]。在大多数情况下，不存在数学上最优的权衡。然而，图8展示了一个关于准确性与对抗鲁棒性之间权衡的假设示例。可行域中任何不在帕累托前沿上的点都是坏点(即帕累托低效)。存在一个更好的解决方案(即帕累托改进)，它可以在不损害另一个目标的情况下显著帮助一个目标，这是帕累托优化的目标。此外，如果在某个用例中存在单一最优解，帕累托优化自然能够达到它。组织可能需要接受这些属性之间的权衡，并根据人工智能系统、用例以及人工智能技术的其他相关影响来决定优先考虑哪些属性[274,326,382]。

4.1.2. 对抗鲁棒性的理论局限性

鉴于存在大量强大的攻击，必须在人工智能系统部署到关键领域之前设计适当的缓解措施。由于该领域许多任务缺乏理论上安全的机器学习算法(见第 1 节)，这一挑战更加严峻。这意味着设计缓解措施本质上是一个临时且易出错的过程，尽管已有保护机器学习部署的实践指南[69,274]和缓解对抗性机器学习攻击的现有指南[120]。

AML中一个持续存在的挑战是检测模型何时受到攻击的能力。了解这一点将提供在信息丢失或模型中触发不良行为之前反击攻击的机会。然而，Tramèr[373]已经表明，设计检测对抗样本的技术等同于鲁棒分类，而这本质上是难以解决的。对抗样本可能来自模型训练时所使用的、且期望输入所属的同一数据分布，也可能是分布外(OOD)输入。因此，检测OOD输入的能力也是AML中的一个重要挑战。Fang等人[124]建立了关于可检测性的有用理论界限，特别是在分布内数据与OOD数据之间存在重叠时的不可能性结果。

形式化方法验证在航空电子和密码学等其他需要高保障的领域有着悠久的历史。尽管应用这种方法论的结果提供了安全性和安全性保障，但其成本非常高，这阻碍了形式化方法的广泛采用。目前，这些领域中的形式化方法主要用于法规强制要求的应用中。将形式化方法应用于神经网络有可能提供急需的安全保障，尤其是在高风险应用中。然而，这项技术的可行性将由技术和业务标准的组合来决定--即以可接受的成本处理当今感兴趣的复杂ML模型的能力。需要更多的研究来将这项技术扩展到ML算法中使用的代数运算，将其扩展到当今使用的大型模型，并适应人工智能系统代码的快速变化，同时限制应用形式化验证的成本。

4.1.3. Evaluation

Another general problem of AML mitigations for both evasion and poisoning attacks is the lack of reliable benchmarks, which causes results from AML papers to be routinely incomparable, as they do not rely on the same assumptions and methods. While there have been some promising developments in this direction [97, 327], more research and encouragement are needed to foster the creation of standardized benchmarks to gain reliable insights into the actual performance of proposed mitigations.

More broadly, the effectiveness of a mitigation is determined not just by how well it will defend against existing attack, but also how well it defends against unforeseen attacks. This means that new mitigations should be tested adversarially, with the researchers proposing the mitigation also trying to break it. This is often difficult and time-consuming, leading to less rigorous and reliable evaluations of novel mitigations; often they appear very powerful, but are quickly shown lack robustness to unforeseen types of attacks.

Finally, this difficulty combines with the difficulty of trading off between different attributes discussed above. Instead of evaluating each attribute in isolation, they should be evaluated simultaneously for any new mitigations, and mitigations should be compared on a Pareto plot (as in Fig. 8) capturing the various tradeoffs that have to be made. This additionally increases the cost to evaluating new mitigations, and can make comparing mitigations difficult - if the green dot represents a new method, it is not possible to say it is an improvement on the red dot, as it is better on one axis but worse on the other.

4.2. Discussion

4.2.1. The Scale Challenge

Data is fundamentally important for training models. Recent trends in GenAI have been towards significant investment in larger models and larger datasets for training them. Few developers of foundation models publish key details about the data sources used in their training [44]. Those who do [247, 371] show the scale of the footprint and the massive amount of data consumed during training. The most recent multi-modal GenAI systems further exacerbate the demand by requiring large amounts of data for each modality.

In most cases, no single entity controls all of the data used to train a particular foundation model. Data repositories are not monolithic data containers but a list of labels and data links to other servers that actually contain the corresponding data samples. This paradigm challenges the classic definition of the corporate cybersecurity perimeter and creates new risks that are difficult to mitigate [57]. Recently published open-source data poisoning tools [241] increase the risk of large-scale attacks on image training data. Although created to enable artists to protect the copyright of their work, these tools may become harmful in the hands of people with malicious intent.

There are several ways this new class of attacks could be mitigated, although it is unclear

4.1.3. 评估

针对规避和投毒攻击的对抗性机器学习缓解措施的另一个普遍问题是缺乏可靠的基准，这导致对抗性机器学习论文的结果通常无法比较，因为它们不依赖于相同的假设和方法。虽然在这一方向上已有一些有前景的进展[97,327]，但仍需要更多研究和鼓励以促进标准化基准的创建，从而获得对所提出缓解措施实际性能的可靠洞察。

更广泛地说，缓解措施的有效性不仅取决于它抵御现有攻击的能力，还取决于它抵御不可预见攻击的能力。这意味着新的缓解措施应该以对抗方式进行测试，提出缓解措施的研究人员也应尝试攻破它。这通常是困难且耗时的，导致对新缓解措施的评估不够严格和可靠；通常它们看起来非常强大，但很快就被证明对不可预见的攻击类型缺乏鲁棒性。

最后，这一困难与上述在不同属性之间进行权衡的困难相结合。不应孤立地评估每个属性，而应同时评估任何新的缓解措施，并且应在帕累托图(如图8)上比较缓解措施，以捕捉必须做出的各种权衡。这进一步增加了评估新缓解措施的成本，并可能使比较缓解措施变得困难--如果绿点代表一种新方法，则无法说它是对红点的改进，因为它在一个轴上更好但在另一个轴上更差。

4.2. 讨论

4.2.1. 规模挑战

数据对于训练模型具有根本性的重要性。近期 GenAI 的趋势是大量投资于更大的模型和用于训练它们的更大数据集。很少有基础模型的开发者公布其训练所用数据源的关键细节[44]。那些公布了的[247,371]则展示了足迹的规模以及训练期间消耗的海量数据。最新的多模态 GenAI 系统进一步加剧了这一需求，因为每种模态都需要大量数据。

在大多数情况下，没有单一实体控制用于训练特定基础模型的全部数据。数据存储库并非单一的数据容器，而是标签列表和指向实际包含相应数据样本的其他服务器的数据链接。这一范式挑战了企业网络安全边界的经典定义，并产生了难以缓解的新风险[57]。近期发布的开源数据投毒工具[241]增加了对图像训练数据进行大规模攻击的风险。尽管这些工具是为了让艺术家保护其作品的版权而创建的，但它们在恶意者手中可能变得有害。

尽管尚不清楚，但有几种方法可以缓解这类新型攻击

how effective mitigations will prove to be as the sophistication of attacks increase. Data and model sanitization techniques (see Sec. 2.3) reduce the impacts of a range of poisoning attacks. They can be combined with cryptographic techniques for origin and integrity attestation to provide assurances downstream, as recommended in the final report of the National Security Commission on AI [267]. Robust training techniques (see Sec. 2.3) offer different approaches to developing theoretically certified defenses against data poisoning attacks with the intention of providing much-needed information-theoretic guarantees for security. The results are encouraging, but more research is needed to extend this methodology to more general assumptions about data distributions, the ability to handle out-of-distribution inputs, more complex models, multiple data modalities, and better performance. Another challenge is applying these techniques to very large models like LLMs and generative diffusion models, which are becoming targets of attacks [55, 90].

4.2.2. Supply Chain Challenges

The literature on AML shows a trend of designing new attacks that are more difficult to detect. Since the poisoning of AI models can persist through safety training and be triggered by attackers on demand [170], significant concerns arise regarding the potential for models to be created with intentional exploits that are hard for organizations deploying and using models to detect. The potential for attacks against open-source dependencies may be particularly acute in the AI context because organizations and researchers may not be able to audit and identify vulnerabilities encoded into a model's weights in the same way it is often possible to audit open-source software. As users come to rely more on the outputs of AI systems — for example, some research suggests that software engineers who over-rely on AI coding assistants' suggestions may produce less secure code [290, 294, 308] — the potential for malicious actors to subtly manipulate the outputs of AI systems may create increased risk.

Additionally, Goldwasser et al. [142] introduced a new class of attacks: information-theoretically undetectable Trojans that can be planted in ML models. If proven practical, the undetectable nature of such attacks would pose significant challenges for AI supply-chain risk management and increase the importance of preventing insider threat throughout the supply chain. DARPA and NIST have also jointly created TrojAI to research the defense of AI systems from intentional, malicious Trojans by developing the technology to detect and investigate these attacks.

4.2.3. Multimodal Models

MULTIMODAL MODELS have shown great potential for achieving high performance on many ML tasks [27, 30, 258, 304, 435]. However, emerging evidence from practice shows that a redundancy of information across the different modalities does not necessarily make the model more robust against adversarial perturbations of a single modality. Combining modalities and training the model on clean data alone does not seem to improve adver-

随着攻击复杂度的提高，缓解措施将被证明有多有效。数据和模型净化技术(见第 2.3 节)可减轻一系列投毒攻击的影响。如美国人工智能国家安全委员会最终报告所建议[267]，它们可以与用于来源和完整性证明的密码学技术相结合，以在下游提供保证。鲁棒训练技术(见第 2.3 节)提供了不同的方法来开发针对数据投毒攻击的理论上可认证的防御，旨在为安全性提供急需的信息论保证。结果令人鼓舞，但还需要更多研究来将这一方法扩展到关于数据分布的更一般假设、处理分布外输入的能力、更复杂的模型、多种数据模态以及更好的性能。另一个挑战是将这些技术应用于 LLM 和生成式扩散模型等超大型模型，这些模型正成为攻击目标[55,90]。

4.2.2. 供应链挑战

关于 AML 的文献显示出一种设计更难检测的新攻击的趋势。由于人工智能模型的投毒可以持续通过安全训练，并由攻击者按需触发[170]，因此人们严重担忧模型可能被创建时带有故意植入的漏洞利用，而部署和使用模型的组织难以检测到。针对开源依赖项的潜在攻击在人工智能背景下可能尤为严重，因为组织和研究人员可能无法像通常审计开源软件那样审计和识别编码到模型权重中的漏洞。随着用户越来越依赖人工智能系统的输出--例如，一些研究表明，过度依赖人工智能编码助手建议的软件工程师可能会产生安全性较低的代码[290,294,308]--恶意行为者微妙操纵人工智能系统输出的可能性可能会带来更大的风险。

此外，Goldwasser 等人[142]引入了一类新的攻击：可植入 ML 模型中的信息论上不可检测的木马。如果被证明可行，此类攻击的不可检测性将对人工智能供应链风险管理构成重大挑战，并增加在整个供应链中防止内部威胁的重要性。DARPA 和 NIST 还联合创建了 TrojAI，通过开发检测和调查这些攻击的技术来研究防御人工智能系统免受故意恶意木马攻击。

4.2.3. 多模态模型

多模态模型在许多机器学习任务上展现出实现高性能的巨大潜力[27,30,258,304,435]。然而，来自实践的新证据表明，不同模态之间的信息冗余并不一定能使模型对单一模态的对抗性扰动更具鲁棒性。结合多种模态并仅在干净数据上训练模型似乎并不能提高对抗

sarial robustness. In addition, adversarial training, which is widely used in single modality applications, may become prohibitively expensive as the number of modality combinations increases. Additional effort is required to benefit from the redundant information in order to improve robustness against single modality attacks [417]. Without such an effort, single modality attacks can be effective and compromise multimodal models across a wide range of multimodal tasks despite the information contained in the remaining unperturbed modalities [417, 424]. Moreover, researchers have devised efficient mechanisms for constructing simultaneous attacks on multiple modalities, which suggests that multimodal models might not be more robust against adversarial attacks despite improved performance [77, 333, 415].

The existence of simultaneous attacks on multimodal models suggests that mitigation techniques that only rely on single modality perturbations are not likely to be robust.

4.2.4. Quantized Models

Quantization is a technique for efficiently deploying models to edge platforms, such as smart phones and IoT devices [138]. It reduces the computational and memory costs of running inference on a given platform by representing the model weights and activations with low-precision data types. For example, quantized models typically use 8-bit integers (int8) or even more compact 4-bit representations instead of the usual 32-bit floating point (float32) numbers for the original non-quantized model.

This technique has been widely used with PredAI and increasingly with GenAI models [108]. However, quantized models inherit the vulnerabilities of the original models and introduce additional weaknesses that make them vulnerable to adversarial attacks. Error amplification from reduced computational precision adversely affects the adversarial robustness of the quantized models. While there are some useful mitigation techniques for PredAI models [216], the effects of quantization on GenAI models have not been studied as thoroughly. Organizations that deploy such models should continuously monitor their behavior. Recent results [118] reveal that widely used quantization methods can be exploited to produce a harmful quantized LLM, even though the full-precision counterpart appears benign, potentially tricking users into deploying the malicious quantized model.

4.2.5. Risk Management in Light of AML

A key question that this taxonomy deliberately leaves aside is how organizations can make decisions about the development and use of AI systems in light of evidence about the increasing diversity of AML attacks and the efficacy and limitations of available mitigations.

Especially in GenAI, some model developers and application builders have moved towards paradigms for testing adversarial risks as part of pre-deployment testing and evaluation

对抗鲁棒性。此外，在单模态应用中广泛使用的对抗训练，随着模态组合数量的增加，可能会变得极其昂贵。需要额外的努力来利用冗余信息，以提高对单模态攻击的鲁棒性[417]。如果没有这样的努力，尽管其余未受扰动的模态中包含信息，单模态攻击仍然可以有效并危害多模态模型，跨越广泛的多模态任务[417,424]。此外，研究人员已经设计了有效的机制来构建对多个模态的同时攻击，这表明多模态模型尽管性能有所提高，但可能并不对对抗攻击更具鲁棒性[77,333,415]。

针对多模态模型的同时攻击的存在表明，仅依赖单一模态扰动的缓解技术不太可能具有鲁棒性。

4.2.4. 量化模型

量化是一种将模型高效部署到边缘平台(如智能手机和物联网设备)的技术[138]。它通过使用低精度数据类型表示模型权重和激活值，降低了在给定平台上运行推理的计算和内存成本。例如，量化模型通常使用8位整数(int8)甚至更紧凑的4位表示，而非原始未量化模型通常使用的32位浮点数(float32)。

这种技术已广泛用于PredAI，并越来越多地用于GenAI模型[108]。然而，量化模型继承了原始模型的漏洞，并引入了额外的弱点，使其容易受到对抗攻击。计算精度降低导致的误差放大会对量化模型的对抗鲁棒性产生不利影响。虽然PredAI模型有一些有用的缓解技术[216]，但量化对GenAI模型的影响尚未得到如此彻底的研究。部署此类模型的组织应持续监控其行为。最近的结果[118]揭示，广泛使用的量化方法可以被利用来产生有害的量化LLM，即使全精度对应模型看起来是良性的，也可能诱使用户部署恶意的量化模型。

4.2.5. 对抗性机器学习视角下的风险管理

本分类法有意搁置的一个关键问题是，面对对抗性机器学习攻击日益多样化以及现有缓解措施的有效性和局限性的证据，组织如何就人工智能系统的开发和使用做出决策。

特别是在生成式人工智能中，一些模型开发者和应用构建者已转向将对抗性风险评估作为部署前测试和评估的一部分的范式

of models, such as through a structured process for RED TEAMING [68, 133]. NIST has produced an initial public draft of guidance for model developers on managing risks associated with the misuse of foundation model capabilities, including through pre-deployment evaluations. [275] NIST [273] and Barrett et al. [28] have also developed risk profiles for generative AI systems that map to the NIST AI RMF [274] that may assist model developers and users in assessing risks, including those from adversarial attacks.

However, a persistent challenge remains in the fact that many AML mitigations are empirical in nature and lack theoretical or provable guarantees. In fact, several research results have pointed to theoretical *limits* on AML mitigations, including the impossibility of model-based detection to prevent all impermissible outputs [140] and findings that, so long as a model has any probability of exhibiting an undesired behavior, there exist prompts that can trigger that behavior, implying that any alignment process that attenuates but does not remove an unwanted behavior will remain vulnerable to adversarial prompting attacks. [406]

These theoretical limitations do not obviate the utility of pre-deployment adversarial testing, since such testing can potentially foreclose many attack vectors and thus increase the difficulty of mounting a successful attack above would-be attackers' threshold of effort or capability. However, they suggest that organizations seeking to manage risks related to the development of models with potentially harmful capabilities or the delegation of trust to models in high stakes contexts may need to consider practices and measures beyond adversarial testing to manage the risks associated with AML attacks.

4.2.6. AML and Other AI System Characteristics

A final consideration with respect to adversarial machine learning, and one closely related to questions of risk management, is how to relate and integrate consideration of AML attacks to definitions and processes relating to other desired AI system characteristics.

For example, managing the security of AI systems will require combining mitigations from the field of AML with best practices for the development of secure software from the field of cybersecurity. Understanding and relating these practices to each other, as well as identifying whether there are other key considerations for AI security that fall outside of the scope of either AML or cybersecurity, will be critical as organizations seek to extend existing cybersecurity processes and best practices to address the security of newly adopted AI systems.

Similarly, robustness to AML attacks may play an important role in areas beyond the remit of security, such as in AI safety [275] or in achieving other characteristics of trustworthy AI systems [274]. AML is neither a complete solution to, nor a subset of, any one of these characteristics, and as such, more precisely relating AML attacks and mitigations to processes for achieving these goals and managing risks in AI systems is an area for ongoing work.

模型，例如通过结构化的红队测试流程[68,133]。NIST已为模型开发者制定了关于管理基础模型能力滥用相关风险的指南的初始公开草案，包括通过部署前评估。[275] NIST[273]和Barrett等人[28]还制定了生成式人工智能系统的风险概况，映射到NIST AI RMF[274]，这可能有助于模型开发者和用户评估风险，包括来自对抗攻击的风险。

然而，一个持续存在的挑战是，许多AML缓解措施本质上是经验性的，缺乏理论或可证明的保证。事实上，一些研究结果指出了AML缓解措施的理论限制，包括基于模型的检测无法防止所有不允许的输出[140]，以及发现只要模型有任何概率表现出不良行为，就存在可以触发该行为的提示，这意味着任何减弱但未消除不良行为的对齐过程仍将容易受到对抗提示攻击。[406]

这些理论限制并不排除部署前对抗测试的效用，因为此类测试可能潜在地排除许多攻击向量，从而将成功攻击的难度提高到潜在攻击者的努力或能力阈值以上。然而，它们表明，寻求管理与具有潜在有害能力的模型开发相关的风险或在高风险情境中将信任委托给模型的组织，可能需要考虑对抗测试之外的实践和措施来管理与AML攻击相关的风险。

4.2.6. 对抗性机器学习(AML)及其他 AI 系统特性

关于对抗性机器学习的最后一个考量，也是与风险管理问题密切相关的考量，是如何将对对抗性机器学习攻击的考量与其他期望的人工智能系统特性相关的定义和流程相关联和整合。

例如，管理人工智能系统的安全需要将对抗性机器学习领域的缓解措施与网络安全领域的安全软件开发最佳实践相结合。理解这些实践并将它们相互关联，以及识别是否存在超出对抗性机器学习或网络安全范围的其它人工智能安全关键考量因素，对于组织寻求扩展现有网络安全流程和最佳实践以应对新采用的人工智能系统的安全问题至关重要。

同样，对对抗性机器学习攻击的鲁棒性可能在安全领域之外发挥重要作用，例如在人工智能安全[275]或实现可信人工智能系统的其他特性[274]方面。对抗性机器学习既不是这些特性中任何一个的完整解决方案，也不是其子集，因此，更精确地将对抗性机器学习攻击和缓解措施与实现这些目标和在人工智能系统中管理风险的流程相关联，是一个持续工作的领域。

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *ACM Conference on Computer and Communications Security*, CCS '16, pages 308–318, 2016. https://arxiv.org/abs/1607.00133. doi:10.48550/arXiv.1607.00133.

[2] Mark Abspoel, Daniel Escudero, and Nikolaj Volgushev. Secure training of decision trees with continuous attributes. In *Proceedings on Privacy Enhancing Technologies (PoPETs) 2021, Issue 1*, 2020.

[3] Hojjat Aghakhani, Wei Dai, Andre Manoel, Xavier Fernandes, Anant Kharkar, Christopher Kruegel, Giovanni Vigna, David Evans, Ben Zorn, and Robert Sim. TrojanPuzzle: Covertly poisoning code-suggestion models, 2024. URL: https://arxiv.org/abs/2301.02344, arXiv:2301.02344, doi:10.48550/arXiv.2301.02344.

[4] Hojjat Aghakhani, Dongyu Meng, Yu-Xiang Wang, Christopher Kruegel, and Giovanni Vigna. Bullseye polytope: A scalable clean-label poisoning attack with improved transferability. In *IEEE European Symposium on Security and Privacy, 2021, Vienna, Austria, September 6-10, 2021*, pages 159–178. IEEE, 2021. URL: https://ieeexplore.ieee.org/document/9581207, doi:10.1109/EuroSP51992.2021.00021.

[5] Meta AI. Llama guard 3 documentation, 2024. Accessed: 2024-08-13. URL: https://llama.meta.com/docs/model-cards-and-prompt-formats/llama-guard-3/.

[6] Meta AI. Prompt guard documentation, 2024. Accessed: 2024-08-13. URL: https://llama.meta.com/docs/model-cards-and-prompt-formats/prompt-guard/.

[7] AI Safety Institute. Systemic ai safety fast grants. https://www.aisi.gov.uk/grants, 2024. Accessed: 2024-08-22.

[8] Dan Alistarh, Zeyuan Allen-Zhu, and Jerry Li. Byzantine Stochastic Gradient Descent. In *NeurIPS*, 2018. URL: https://arxiv.org/abs/1803.08917, doi:10.48550/arXiv.1803.08917.

[9] Gabriel Alon and Michael Kamfonas. Detecting language model attacks with perplexity, 2023. URL: https://arxiv.org/abs/2308.14132, arXiv:2308.14132, doi:10.48550/arXiv.2308.14132.

[10] Galen Andrew, Peter Kairouz, Sewoong Oh, Alina Oprea, H. Brendan McMahan, and Vinith Suriyakumar. One-shot empirical privacy estimation for federated learning, 2023. URL: https://arxiv.org/abs/2302.03098, arXiv:2302.03098, doi:10.48550/arXiv.2302.03098.

[11] Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. Jailbreaking leading safety-aligned LLMs with simple adaptive attacks, 2024. URL: https://arxiv.org/abs/2404.02151, arXiv:2404.02151, doi:10.48550/arXiv.2404.02151.

[12] Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, Zico Kolter, Matt Fredrikson, Eric Winsor, Jerome Wynne, Yarin Gal, and Xander Davies. Agentharm: A benchmark for measuring harmfulness of llm agents, 2024. URL: https://arxiv.org/abs/2410.09024, arXiv:2410.09024.

参考文献

[1] Martin Abadi、Andy Chu、Ian Goodfellow、H Brendan McMahan、Ilya Mironov、Kunal Talwar 和 Li Zhang。带差分隐私的深度学习。载于*ACM* 计算机与通信安全会议,CCS '16，第308-318 页，2016 年。https: //arxiv.org/abs/1607.00133。doi:10.48550/arXiv.1607.00133.

[2] Mark Abspoel、Daniel Escudero 和 Nikolaj Volgushev。具有连续属性的决策树的安全训练。载于《隐私增强技术会议论文集*(PoPETs)*》*2021* 年第 *1* 期,2020 年。

[3] 霍贾特·阿加哈尼、戴伟、安德烈·马诺埃尔、泽维尔·费尔南德斯、阿南特·卡尔卡尔、克里斯托-pher Kruegel、Giovanni Vigna、David Evans、Ben Zorn 和 Robert Sim。TrojanPuzzle：隐蔽地投毒代码建议模型，2024。URL:https://arxiv.org/abs/2301. 02344,arXiv:2301.02344,doi:10.48550/arXiv.2301.02344.

[4] Hojjat Aghakhani、Dongyu Meng、Yu-Xiang Wang、Christopher Kruegel 和 Giovanni Vigna。Bullseye polytope：一种具有改进可迁移性的可扩展干净标签投毒攻击。载于*IEEE* 欧洲安全与隐私研讨会，*2021*，奥地利维也纳，*2021*年*9*月*6*日至*10*日，第159-178页。IEEE,2021。URL:https://ieeexplo re.ieee.org/document/9581207,doi:10.1109/EuroSP51992.2021.00021.

[5] Meta AI。Llama guard 3 文档，2024 年。访问日期：2024-08-13。网址：https: //llama.meta.com/docs/model-cards-and-prompt-formats/llama-guard-3/。

[6] Meta AI。提示防护文档，2024年。访问日期：2024-08-13。网址：https: //llama.meta.com/docs/model-cards-and-prompt-formats/prompt-guard/。

[7] AI 安全研究所。系统性 AI 安全快速资助。https://www.aisi.gov.uk/grants, 2024. 访问日期：2024-08-22。

[8] Dan Alistarh、Zeyuan Allen-Zhu 和 Jerry Li。拜占庭随机梯度下降。载于*NeurIPS*,2018 年。URL:https://arxiv.org/abs/1803.08917, doi:10.48550/arXiv.1803.08917。

[9] 加布里埃尔·阿隆和迈克尔·卡姆福纳斯。使用逐检测语言模型攻击 plexity，2023年。网址：https://arxiv.org/abs/2308.14132, arXiv:2308.14132，doi:10.48550/arXiv.2308.14132。

[10] Galen Andrew、Peter Kairouz、Sewoong Oh、Alina Oprea、H.Brendan McMahan 和 Vinith Suriyakumar。面向联邦学习的一次性经验隐私估计，2023。URL:https:// arxiv.org/abs/2302.03098, arXiv:2302.03098,doi:10.485 50/arXiv.2302.03098。

[11] Maksym Andriushchenko、Francesco Croce 和 Nicolas Flammarion。用简单的自适应攻击越狱领先的安全对齐 LLM,2024 年。URL:https://arxiv.org/ abs/2404.02151,arXiv:2404.02151,doi:10.48550/arXiv.2404.02151.

[12] Maksym Andriushchenko、Alexandra Souly、Mateusz Dziemian、Derek Duenas、Maxwell Lin、Justin Wang、Dan Hendrycks、Andy Zou、Zico Kolter、Matt Fredrikson、Eric Winsor、Jerome Wynne、Yarin Gal 和 Xander Davies。Agentharm：用于衡量 LLM 智能体危害性的基准，2024 年。网址：https://arxiv.org/abs/24 10.09024,arXiv:2410.09024。

[13] Anthropic. Model Card and Evaluations for Claude Models. https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf, July 2023. Anthropic.

[14] Anthropic. Anthropic's interactive prompt engineering tutorial. https://github.com/anthropics/prompt-eng-interactive-tutorial, 2024. Accessed: 2024-08-22.

[15] Anthropic. Claude 3.5 Sonnet. https://www.anthropic.com/news/claude-3-5-sonnet, June 2024. Anthropic.

[16] Anthropic. Expanding our model safety bug bounty program. https://www.anthropic.com/news/model-safety-bug-bounty, 2024. Accessed: 2024-08-22.

[17] Giovanni Apruzzese, Hyrum S Anderson, Savino Dambra, David Freeman, Fabio Pierazzi, and Kevin Roundy. "real attackers don't compute gradients": Bridging the gap between adversarial ml research and practice. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 339–364. IEEE, 2023. URL: https://arxiv.org/abs/2212.14315, doi:10.48550/arXiv.2212.14315.

[18] Arthur. Shield, 2023. URL: https://www.arthur.ai/product/shield.

[19] Giuseppe Ateniese, Luigi V. Mancini, Angelo Spognardi, Antonio Villani, Domenico Vitali, and Giovanni Felici. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers. *Int. J. Secur. Netw.*, 10(3):137–150, September 2015. doi:10.1504/IJSN.2015.071829.

[20] Anish Athalye, Nicholas Carlini, and David A. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 274–283. PMLR, 2018. URL: http://proceedings.mlr.press/v80/athalye18a.html, doi:10.48550/arXiv.1802.00420.

[21] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples, 2018. URL: https://arxiv.org/abs/1707.07397, arXiv:1707.07397, doi:10.48550/arXiv.1707.07397.

[22] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2938–2948. PMLR, 26–28 Aug 2020. URL: http://proceedings.mlr.press/v108/bagdasaryan20a.html.

[23] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *AISTATS*. PMLR, 2020. URL: https://proceedings.mlr.press/v108/bagdasaryan20a.html, doi:10.48550/arXiv.1807.00459.

[24] Eugene Bagdasaryan, Ren Yi, Sahra Ghalebikesabi, Peter Kairouz, Marco Gruteser, Sewoong Oh, Borja Balle, and Daniel Ramage. Air gap: Protecting privacy-conscious conversational agents, 2024. URL: https://arxiv.org/abs/2405.05175, arXiv:2405.05175, doi:10.48550/arXiv.2405.05175.

[13] Anthropic。Claude 模型模型卡与评估。https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf，2023年7月。Anthropic。

[14] Anthropic。Anthropic 的交互式提示工程教程。https://github.com/anthropics/prompt-eng-interactive-tutorial,2024。访问日期：2024-08-22。

[15] Anthropic。Claude 3.5 Sonnet。https://www.anthropic.com/news/claude-3-5-sonnet,2024年6月。Anthropic。

[16] Anthropic。扩展我们的模型安全漏洞赏金计划。https://www.anthropic.com/news/model-safety-bug-bounty,2024。访问日期：2024-08-22。

[17] 乔瓦尼·阿普鲁泽塞，海勒姆·S·安德森，萨维诺·丹布拉，大卫·弗里曼，法比奥·皮耶尔-Azzi 和 Kevin Roundy。"真实攻击者不计算梯度"：弥合对抗性机器学习研究与实践之间的差距。载于*2023 IEEE* 安全与可信机器学习会议*(SaTML)*，第 339-364 页。IEEE,2023 年。URL:https://arxiv.org/abs/2212.14315, doi:10.48550/arXiv.2212.14315.

[18] Arthur。Shield，2023年。网址：https://www.arthur.ai/product/shield.

[19] Giuseppe Ateniese, Luigi V.Mancini, Angelo Spognardi, Antonio Villani, Domenico Vitali, and Giovanni Felici.用更聪明的机器入侵智能机器：如何从机器学习分类器中提取有意义的数据。*Int.J.Secur.Netw.*, 10(3): 137-150, September 2015. doi:10.1504/IJSN.2015.071829.

[20] Anish Athalye、Nicholas Carlini 和 David A.Wagner。混淆梯度给人以虚假的安全感：规避对抗样本的防御。载于 Jennifer G.Dy 和 Andreas Krause 编，第*35*届国际机器学习会议论文集，*ICML 2018*，瑞典斯德哥尔摩 *Stockholmsmässan,2018*年7月*10-15*日，第80卷，机器学习研究论文集，第274-283页。PMLR,2018年。URL: http://Proceedings.mlr.press/v80/athalye18a.html, doi:10.48550/arXiv.1802.00420。

[21] [353] Anish Athalye、Logan Engstrom、Andrew Ilyas 和 Kevin Kwok。合成鲁棒对抗样本，2018。网址：⟨https://arxiv.org/abs/1707.07397,⟩ arXiv:1707.07397,doi:⟨10.48550/arXiv.1707.07397.⟩

[22] Eugene Bagdasaryan、Andreas Veit、Yiqing Hua、Deborah Estrin 和 Vitaly Shmatikov。如何后门联邦学习。载于 Silvia Chiappa 和 Roberto Calandra 编，第二十三届国际人工智能与统计学会议论文集，第 108 卷，机器学习研究论文集，第 2938-2948 页。PMLR,2020 年 8 月 26-28 日。网址：http://Proceedings.mlr.press/v108/Bagdasaryan20a.html。

[23] Eugene Bagdasaryan、Andreas Veit、Yiqing Hua、Deborah Estrin 和 Vitaly Shmatikov。如何后门联邦学习。载于*AISTATS*。PMLR,2020。URL:https://proceedings.mlr.press/v108/Bagdasaryan20a.html, doi:10.48550/arXiv.1807.00459。

[24] Eugene Bagdasaryan, Ren Yi, Sahra Ghalebikesabi, Peter Kairouz, Marco Gruteser, Sewoong Oh, Borja Balle， 和 Daniel Ramage。空气隔离：保护注重隐私的对话代理，2024年。网址：https://arxiv.org/abs/2405.05175, arXiv: 2405.05175, doi:10.48550/arXiv.2405.05175.

[25] Marieke Bak, Vince Istvan Madai, Marie-Christine Fritzsche, Michaela Th. Mayrhofer, and Stuart McLennan. You can't have ai both ways: Balancing health data privacy and access fairly. *Frontiers in Genetics*, 13, 2022. https://www.frontiersin.org/articles/10.3389/fgene.2022.929453. doi:10.3389/fgene.2022.929453.

[26] Borja Balle, Giovanni Cherubin, and Jamie Hayes. Reconstructing training data with informed adversaries. In *NeurIPS 2021 Workshop on Privacy in Machine Learning (PRIML)*, 2021. URL: https://openreview.net/forum?id=Yi2DZTbnBI4, doi:10.48550/arXiv.2201.04845.

[27] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy, 2017. doi:10.48550/ARXIV.1705.09406.

[28] Anthony M. Barrett, Dan Hendrycks, Jessica Newman, and Brandie Nonnecke. *UC Berkeley AI Risk-Management Standards Profile for General-Purpose AI Systems (GPAIS) and Foundation Models*. UC Berkeley Center for Long Term Cybersecurity, 2023. https://cltc.berkeley.edu/seeking-input-and-feedback-ai-risk-managem ent-standards-profile-for-increasingly-multi-purpose-or-general-purpose-ai/. doi:10.48550/ARXIV.2206.08966.

[29] Lejla Batina, Shivam Bhasin, Dirmanto Jap, and Stjepan Picek. CSI NN: Reverse engineering of neural network architectures through electromagnetic side channel. In *Proceedings of the 28th USENIX Conference on Security Symposium*, SEC'19, page 515–532, USA, 2019. USENIX Association. URL: https://www.usenix.org/conference/usenixsecurity19/presentation/batina.

[30] Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa. A survey on deep multimodal learning for computer vision: Advances, trends, applications, and datasets. *Vis. Comput.*, 38(8):2939–2970, August 2022. doi:10.1007/s00371-021-02166-7.

[31] Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023. URL: https://arxiv.org/abs/2303.08112, doi:10.48550/arXiv.2303.08112.

[32] Philipp Benz, Chaoning Zhang, Soomin Ham, Gyusang Karjauv, Adil Cho, and In So Kweon. The triangular trade-off between accuracy, robustness, and fairness. *Workshop on Adversarial Machine Learning in Real-World Computer Vision Systems and Online Challenges (AML-CV) at CVPR*, 2021. URL: https://dl.acm.org/doi/10.1145/3645088, doi:10.1145/3645088.

[33] Jamie Bernardi, Gabriel Mukobi, Hilary Greaves, Lennart Heim, and Markus Anderljung. Societal adaptation to advanced ai, 2024. URL: https://arxiv.org/abs/2405.10295, arXiv:2405.10295, doi:10.48550/arXiv.2405.10295.

[34] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 5050–5060.

[25] 玛丽克·巴克，文斯·伊什特万·马达伊，玛丽-克里斯汀·弗里切，米夏埃拉·Th. Mayrhofer, and Stuart McLennan.你不能两全其美地对待 AI：公平地平衡健康数据隐私与访问。*Frontiers in Genetics*, 13, 2022. https://www.fronti ersin.org/articles/10.3389/fgene.2022.929453. doi:10.3389/fgene.2022.929 453.

[26] Borja Balle, Giovanni Cherubin， 和 Jamie Hayes。利用知情对手重建训练数据。载于 NeurIPS 2021 机器学习隐私研讨会 (PRIML)，2021年。网址：https://openreview.net/forum?id=Yi2DZTbnBI4, doi: 10.48550/arXiv.2201.04845.

[27] Tadas Baltrušaitis, Chaitanya Ahuja， 和 Louis-Philippe Morency。多模态机器学习：综述与分类法，2017年。doi:10.48550/ARXIV.1705.09406.

[28] 安东尼·M·巴雷特、丹·亨德里克斯、杰西卡·纽曼和布兰迪·诺内克。加州大学伯克利分校通用人工智能系统*(GPAIS)*与基础模型的人工智能风险管理标准概况。加州大学伯克利分校长期网络安全中心，2023年。https://cltc.berkeley.edu/seeking-input-and-feedback-ai-risk-managem ent-standards-profile-for-increasingly-multi-purpose-or-general-purpose-ai/。

doi:10.48550/ARXIV.2206.08966。

[29] Lejla Batina, Shivam Bhasin, Dirmanto Jap， 和 Stjepan Picek。CSI NN：逆向工通过电磁侧信道逆向工程神经网络架构。载于第*28*届*USENIX*安全研讨会论文集,SEC'19，第515-532页，美国，2019年。USENIX协会。网址：https://www.usenix.org/conferenc e/usenixsecurity19/presentation/batina。

[30] Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui， 和 Abdellatif Mtibaa。计算机视觉深度多模态学习综述：进展、趋势、应用和数据集。Vis. Comput., 38(8):2939-2970，2022年8月。doi:10.1007/s00371-0 21-02166-7。

[31] 诺拉·贝尔罗斯、扎克·弗曼、洛根·史密斯、丹尼·哈拉维、伊戈尔·奥斯特洛夫斯基、列夫·麦肯齐 inney, Stella Biderman， 和 Jacob Steinhardt。利用调谐透镜从变换器中引出潜在预测。arXiv 预印本 arXiv:2303.08112，2023年。网址：https://arxiv.org/abs/2303.08112, doi:10.48550/arXiv.2303.08112.

[32] Philipp Benz、Chaoning Zhang、Soomin Ham、Gyusang Karjauv、Adil Cho 和 In So Kweon。准确率、鲁棒性与公平性之间的三角权衡。*CVPR* 真实世界计算机视觉系统中的对抗性机器学习与在线挑战研讨会*(AML-CV)*,2021。URL:https://dl.acm.org/doi/10.1145/3 645088,doi:10.1145/3645088.

[33] 杰米·贝尔纳迪、加布里埃尔·穆科比、希拉里·格里夫斯、伦纳特·海姆和马库斯·安derljung。社会对先进 AI 的适应，2024。URL:https://arxiv.org/abs/2405.10295,arXiv:2405.10295,doi:10.48550/arXiv.2405.10295.

[34] David Berthelot、Nicholas Carlini、Ian Goodfellow、Nicolas Papernot、Avital Oliver 和 Colin A Raffel。Mixmatch：半监督学习的整体方法。载于 H.Wallach、H.Larochelle、A.Beygelzimer、F.d' Alché-Buc、E.Fox 和 R.Garnett 编，*Advances in Neural Information Processing Systems 32* , 第 5050-5060 页。

Curran Associates, Inc., 2019. URL: http://papers.nips.cc/paper/8749-mixmatch-a-holistic-approach-to-semi-supervised-learning.pdf, doi:10.48550/arXiv.2405.10295.

[35] Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal, and Seraphin Calo. Model Poisoning Attacks in Federated Learning. In *NeurIPS SECML*, 2018.

[36] Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal, and Seraphin Calo. Analyzing federated learning through an adversarial lens. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 634–643. PMLR, 09–15 Jun 2019. URL: https://proceedings.mlr.press/v97/bhagoji19a.html, doi:10.48550/arXiv.1811.12470.

[37] Battista Biggio, Igino Corona, Giorgio Fumera, Giorgio Giacinto, and Fabio Roli. Bagging classifiers for fighting poisoning attacks in adversarial classification tasks. In *Proceedings of the 10th International Conference on Multiple Classifier Systems*, MCS'11, page 350–359, Berlin, Heidelberg, 2011. Springer-Verlag. URL: https://api.semanticscholar.org/CorpusID:12680508.

[38] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013. doi:10.1007/978-3-642-40994-3_25.

[39] Battista Biggio, Blaine Nelson, and Pavel Laskov. Support vector machines under adversarial label noise. In Chun-Nan Hsu and Wee Sun Lee, editors, *Proceedings of the Asian Conference on Machine Learning*, volume 20 of *Proceedings of Machine Learning Research*, pages 97–112, South Garden Hotels and Resorts, Taoyuan, Taiwain, 14–15 Nov 2011. PMLR. URL: https://proceedings.mlr.press/v20/biggio11.html, doi:10.48550/arXiv.2206.00352.

[40] Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In *Proceedings of the 29th International Coference on International Conference on Machine Learning, ICML*, 2012. URL: https://arxiv.org/abs/1206.6389, doi:10.48550/arXiv.1206.6389.

[41] Battista Biggio, Konrad Rieck, Davide Ariu, Christian Wressnegger, Igino Corona, Giorgio Giacinto, and Fabio Roli. Poisoning behavioral malware clustering. In *Proceedings of the 2014 Workshop on Artificial Intelligent and Security Workshop*, AISec '14, page 27–36, New York, NY, USA, 2014. Association for Computing Machinery. doi:10.1145/2666652.2666666.

[42] Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, December 2018. URL: https://doi.org/10.1016%2Fj.patcog.2018.07.023, doi:10.1016/j.patcog.2018.07.023.

[43] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent. In *NeurIPS*, 2017.

Curran Associates, Inc., 2019. 网址：http://papers.nips.cc/paper/8749-mixmatch-a -holistic-approach-to-semi-supervised-learning.pdf，doi:10.48550/arXiv.240 5.10295。

[35] Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal，和 Seraphin Calo。联邦学习中的模型投毒攻击。载于 NeurIPS SECML，2018年。

[36] Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal，和 Seraphin Calo。通-从对抗性视角分析联邦学习。载于 Kamalika Chaudhuri 和 Ruslan Salakhutdinov 编，《第36届国际机器学习会议论文集》，《机器学习研究论文集》第97卷，第634-643页。PMLR，2019年6月9-15日。URL: https://proceedings.mlr.press/v97/bhagoji i19a.html，doi:10.48550/arXiv.1811.12470.

[37] Battista Biggio, Igino Corona, Giorgio Fumera, Giorgio Giacinto，和 Fabio Roli。Bag-用于对抗分类任务中对抗投毒攻击的 ging 分类器。载于第*10*届多分类器系统国际会议论文集,MCS'11，第350-359页，柏林、海德堡，2011年。Springer-Verlag。URL:https://api.semanticscholar.org/CorpusID:12680508。

[38] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto，和 Fabio Roli。测试时针对机器学习的规避攻击。载于机器学习与知识发现数据库联合欧洲会议，第387-402页。Springer，2013年。doi:10.1007/978-3-642-409 94-3_25。

[39] Battista Biggio, Blaine Nelson，和 Pavel Laskov。对抗条件下的支持向量机-对抗性标签噪声。载于 Chun-Nan Hsu 与 Wee Sun Lee 编，*Proceedings of the Asian Conference on Machine Learning*，第 20 卷，*Proceedings of Machine Learning Research*，第 97-112 页，South Garden Hotels and Resorts，台湾桃园，2011 年 11 月 14-15 日。PMLR。URL: https://Proceedings.mlr.press/v20/biggio11.html,

[40] Battista Biggio、Blaine Nelson 和 Pavel Laskov。针对支持向量机的投毒(poisoning)攻击。载于第*29*届国际机器学习大会论文集，*ICML*,2012年。URL:https://arxiv.org/abs/1206.638 9,doi:10.48550/arXiv.1206.6389.

[41] Battista Biggio, Konrad Rieck, Davide Ariu, Christian Wressnegger, Igino Corona, Giorgio Giacinto，和 Fabio Roli。投毒行为恶意软件聚类。载于2014年人工智能与安全研讨会论文集，AISec '14，第27–36页，美国纽约州纽约市，2014年。美国计算机协会。doi:10.1145/2666652.2666666。

[42] Battista Biggio 和 Fabio Roli。野性模式：对抗性机器学习兴起十年之后。Pattern Recognition，84:317-331，2018年12月。URL：https: // doi.org/10.1016%2Fj.patcog.2018.07.023，doi:10.1016/j.patcog.2018.07.0 23.

[43] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui，和 Julien Stainer。与对手共存的机器学习：拜占庭容错梯度下降。载于 NeurIPS，2017年。

URL: https://papers.nips.cc/paper_files/paper/2017/file/f4b9ec30ad9f68f89b29639786cb62ef-Paper.pdf.

[44] Rishi Bommasani, Kevin Klyman, Sayash Kapoor, Shayne Longpre, Betty Xiong, Nestor Maslej, and Percy Liang. The foundation model transparency index v1.1: May 2024, 2024. URL: https://arxiv.org/abs/2407.12929, arXiv:2407.12929, doi:10.48550/arXiv.2407.12929.

[45] Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *42nd IEEE Symposium on Security and Privacy, SP 2021, San Francisco, CA, USA, 24-27 May 2021*, pages 141–159. IEEE, 2021. doi:10.1109/SP40001.2021.00019.

[46] Dillon Bowen, Brendan Murphy, Will Cai, David Khachaturov, Adam Gleave, and Kellin Pelrine. Scaling laws for data poisoning in llms, 2024. URL: https://arxiv.org/abs/2408.02946, arXiv:2408.02946, doi:10.48550/arXiv.2408.02946.

[47] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL: https://openreview.net/forum?id=SyZI0GWCZ, doi:10.48550/arXiv.1712.04248.

[48] Gavin Brown, Mark Bun, Vitaly Feldman, Adam Smith, and Kunal Talwar. When is memorization of irrelevant training data necessary for high-accuracy learning? In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2021, page 123–132, New York, NY, USA, 2021. Association for Computing Machinery. doi:10.1145/3406325.3451131.

[49] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. URL: https://arxiv.org/abs/2005.14165, arXiv:2005.14165.

[50] Gon Buzaglo, Niv Haim, Gilad Yehudai, Gal Vardi, and Michal Irani. Reconstructing training data from multiclass neural networks, 2023. URL: https://arxiv.org/abs/2305.03350, arXiv:2305.03350, doi:10.48550/arXiv.2305.03350.

[51] Xiaoyu Cao, Minghong Fang, Jia Liu, and Neil Zhenqiang Gong. FLTrust: Byzantine-robust federated learning via trust bootstrapping. In *NDSS*, 2021. URL: https://arxiv.org/abs/2012.13995, doi:10.48550/arXiv.2012.13995.

[52] Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*, pages 463–480, 2015. URL: https://ieeexplore.ieee.org/document/7163042, doi:10.1109/SP.2015.35.

网址：https://papers.nips.cc/paper_files/paper/2017/file/f4b9ec30ad9f68f89b29639786cb62ef-Paper.pdf。

[44] Rishi Bommasani、Kevin Klyman、Sayash Kapoor、Shayne Longpre、Betty Xiong、Nestor Maslej 和 Percy Liang。基础模型透明度指数 v1.1：2024年5月，2024年。网址：https://arxiv.org/abs/2407.12929, arXiv:2407.12929，

doi:10.48550/arXiv.2407.12929。

[45] Lucas Bourtoule、Varun Chandrasekaran、Christopher A.Choquette-Choo、Hengrui Jia、Adelin Travers、Baiwu Zhang、David Lie 和 Nicolas Papernot。机器遗忘。载于第*42*届*IEEE*安全与隐私研讨会，*SP 2021*，美国加利福尼亚州旧金山，*2021*年*5*月*24-27*日，第141-159页。IEEE,2021年。doi:10.1109/SP40001.2021. 00019。

[46] Dillon Bowen, Brendan Murphy, Will Cai, David Khachaturov, Adam Gleave， 和 Kellin Pelrine。大型语言模型中数据投毒的缩放定律，2024年。网址：https://arxiv. org/abs/2408.02946, arXiv:2408.02946, doi:10.48550/arXiv.2408.02946.

[47] Wieland Brendel, Jonas Rauber, and Matthias Bethge.基于决策的对抗攻击：针对黑盒机器学习模型的可靠攻击。In *6th Interna- tional Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.OpenReview.net, 2018. URL: https://OpenReview.net/forum?id=SyZI0GWCZ, doi:10.48550/arXiv.1712.04 248.

[48] Gavin Brown, Mark Bun, Vitaly Feldman, Adam Smith, and Kunal Talwar.无关训练数据的记忆何时对高精度学习是必要的?In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2021, page 123-132, New York, NY, USA, 2021. Association for Computing

Machinery。doi:10.1145/3406325.3451131。

[49] Tom B. Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell、Sandhini Agarwal、Ariel Herbert-Voss、Gretchen Krueger、Tom Henighan、Rewon Child、Aditya Ramesh、Daniel M. Ziegler、Jeffrey Wu、Clemens Winter、Christopher Hesse、Mark Chen、Eric Sigler、Mateusz Litwin、Scott Gray、Benjamin Chess、Jack Clark、Christopher Berner、Sam McCandlish、Alec Radford、Ilya Sutskever 和 Dario Amodei。语言模型是少样本学习器。CoRR，abs/2005.14165，2020。网址：https://arxiv.or g/abs/2005.14165，arXiv:2005.14165。

[50] Gon Buzaglo、Niv Haim、Gilad Yehudai、Gal Vardi 和 Michal Irani。从多类神经网络重建训练数据，2023。网址：https://arxiv.org/abs/23 05.03350,arXiv:2305.03350,doi:10.48550/arXiv.2305.03350.

[51] [353] Xiaoyu Cao、Minghong Fang、Jia Liu 和 Neil Zhenqiang Gong。FLTrust：通过信任引导实现拜占庭鲁棒的联邦学习。载于*NDSS*,2021。URL:https://arxi v.org/abs/2012.13995,doi:10.48550/arXiv.2012.13995.

[52] 曹寅智和杨俊峰。迈向通过机器遗忘使系统遗忘--

in。载于*2015 IEEE Symposium on Security and Privacy*，第463-480页，2015年。网址：https://ieeexplore.ieee.org/document/7163042, doi:10.1109/SP.2015.35.

[53] Nicholas Carlini. Poisoning the unlabeled dataset of Semi-Supervised learning. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 1577–1592. USENIX Association, August 2021. URL: https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-poisoning.

[54] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (S&P)*, pages 1519–1519, Los Alamitos, CA, USA, May 2022. IEEE Computer Society. URL: https://doi.ieeecomputersociety.org/10.1109/SP46214.2022.00090, doi:10.1109/SP46214.2022.00090.

[55] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models, 2023. URL: https://arxiv.org/abs/2301.13188, arXiv:2301.13188, doi:10.48550/arXiv.2301.13188.

[56] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. https://arxiv.org/abs/2202.07646, 2022. doi:10.48550/ARXIV.2202.07646.

[57] Nicholas Carlini, Matthew Jagielski, Christopher A Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. Poisoning web-scale training datasets is practical. *arXiv preprint arXiv:2302.10149*, 2023. URL: https://arxiv.org/abs/2302.10149, doi:10.48550/arXiv.2302.10149.

[58] Nicholas Carlini, Matthew Jagielski, and Ilya Mironov. Cryptanalytic extraction of neural network models. In Daniele Micciancio and Thomas Ristenpart, editors, *Advances in Cryptology – CRYPTO 2020*, pages 189–218, Cham, 2020. Springer International Publishing. URL: https://arxiv.org/abs/2003.04884, doi:10.48550/arXiv.2003.04884.

[59] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The Secret Sharer: Evaluating and testing unintended memorization in neural networks. In *USENIX Security Symposium*, USENIX '19), pages 267–284, 2019. https://arxiv.org/abs/1802.08232. URL: https://arxiv.org/abs/1802.08232, doi:10.48550/arXiv.1802.08232.

[60] Nicholas Carlini, Milad Nasr, Christopher A Choquette-Choo, Matthew Jagielski, Irena Gao, Anas Awadalla, Pang Wei Koh, Daphne Ippolito, Katherine Lee, Florian Tramer, et al. Are aligned neural networks adversarially aligned? *arXiv preprint arXiv:2306.15447*, 2023. URL: https://arxiv.org/abs/2306.15447, doi:10.48550/arXiv.2306.15447.

[61] Nicholas Carlini, Daniel Paleka, Krishnamurthy Dj Dvijotham, Thomas Steinke, Jonathan Hayase, A. Feder Cooper, Katherine Lee, Matthew Jagielski, Milad Nasr, Arthur Conmy, Itay Yona, Eric Wallace, David Rolnick, and Florian Tramèr. Stealing part of a production language model, 2024. URL: https://arxiv.org/abs/2403.06634, arXiv:2403.06634, doi:10.48550/arXiv.2403.06634.

[62] Nicholas Carlini, Florian Tramer, Krishnamurthy Dj Dvijotham, Leslie Rice, Mingjie

[53] Nicholas Carlini.对半监督学习的未标记数据集进行投毒。In *30th USENIX Security Symposium (USENIX Security 21)*, pages 1577-1592. USENIX Association, August 2021. URL: https://www.usenix.org/conference/USENIX Securi ty21/presentation/carlini-poisoning.

[54] Nicholas Carlini、Steve Chien、Milad Nasr、Shuang Song、Andreas Terzis 和 Florian Tramer。基于第一原理的成员推断攻击。载于*2022 IEEE Symposium on Security and Privacy (S&P)*，第 1519-1519 页，Los Alamitos, CA, USA,2022 年 5 月。IEEE Computer Society。URL: https://doi.IEEE Computer Society.org/10.1109/SP46 214.2022.00090,doi:10.1109/SP46214.2022.00090.

[55] 尼古拉斯·卡里尼、杰米·海斯、米拉德·纳斯尔、马修·贾吉尔斯基、维卡什·塞瓦格、弗洛里安·特拉梅尔、博尔哈·巴列、达芙妮·伊波利托和埃里克·华莱士。从扩散模型中提取训练数据，2023 年。网址：https://arxiv.org/abs/2301.13188, arXiv:2301.131 88,doi:10.48550/arXiv.2301.13188.

[56] 尼古拉斯·卡里尼、达芙妮·伊波利托、马修·贾吉尔斯基、凯瑟琳·李、弗洛里安·特拉默和迟远·张。量化神经语言模型中的记忆。ht tps://arxiv.org/abs/2202.07646,2022。doi:《10.48550/ARXIV.2202.07646.》

[57] Nicholas Carlini, Matthew Jagielski, Christopher A Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr.对网络规模训练数据集进行投毒是可行的。*arXiv preprint arXiv:2302.10149*, 2023. URL: https://arxiv.org/abs/2302.10149, doi:10.48550/arXiv.2302. 10149.

[58] Nicholas Carlini, Matthew Jagielski, and Ilya Mironov.神经网络模型的密码分析提取。载于 Daniele Micciancio 和 Thomas Ristenpart 编，密码学进展 - *CRYPTO 2020*，第 189-218 页，Cham,2020。Springer International Publishing。URL:https://arxiv.org/abs/2003.04884, doi:10.48550/arX iv.2003.04884。

[59] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song.秘密分享者：评估和测试神经网络中的非预期记忆。载于*USENIX* 安全研讨会,USENIX '19)，第 267-284 页，2019。https://arxiv.org/ abs/1802.08232。URL:https://arxiv.org/abs/1802.08232, doi:10.48550/arXiv.1802.08232。

[60] Nicholas Carlini, Milad Nasr, Christopher A Choquette-Choo, Matthew Jagielski, Irena Gao, Anas Awadalla, Pang Wei Koh, Daphne Ippolito, Katherine Lee, Florian Tramer， 等。对齐的神经网络是否具有对抗性对齐?*arXiv* 预印本 *arXiv:2306.15447*,2023。URL:https://arxiv.org/abs/2306.15447, doi:10.48550/arXiv.2306.15447。

[61] Nicholas Carlini, Daniel Paleka, Krishnamurthy Dj Dvijotham, Thomas Steinke, Jonathan Hayase, A. Feder Cooper, Katherine Lee, Matthew Jagielski, Milad Nasr, Arthur Conmy, Itay Yona, Eric Wallace, David Rolnick， 和 Florian Tramèr。窃取生产语言模型的一部分，2024 年。网址：https://arxiv.org/abs/2403.06634, arXiv:2403.06634, doi:10.48550/arXiv.2403.06634.

[62] 尼古拉斯·卡尔里尼(Nicholas Carlini)、弗洛里安·特拉默(Florian Tramer)、克里希纳穆尔蒂·吉·德维约特哈姆(Krishnamurthy Dj Dvijotham)、莱斯利·赖斯(Leslie Rice)、明杰(Mingjie)

Sun, and J. Zico Kolter. (certified!!) adversarial robustness for free!, 2023. URL: https://arxiv.org/abs/2206.10550, arXiv:2206.10550, doi:10.48550/arXiv.2206.10550.

[63] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association, August 2021. URL: https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

[64] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, AISec '17, page 3–14, New York, NY, USA, 2017. Association for Computing Machinery. doi:10.1145/3128572.3140444.

[65] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Proc. IEEE Security and Privacy Symposium*, 2017. URL: https://arxiv.org/abs/1608.04644, doi:10.48550/arXiv.1608.04644.

[66] Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 1–7. IEEE, 2018. URL: https://arxiv.org/abs/1801.01944, doi:10.48550/arXiv.1801.01944.

[67] Stephen Casper, Yuxiao Li, Jiawei Li, Tong Bu, Kevin Zhang, Kaivalya Hariharan, and Dylan Hadfield-Menell. Red teaming deep neural networks with feature synthesis tools. *arXiv preprint arXiv:2302.10894*, 2023. URL: https://arxiv.org/abs/2302.10894, doi:10.48550/arXiv.2302.10894.

[68] Stephen Casper, Jason Lin, Joe Kwon, Gatlen Culp, and Dylan Hadfield-Menell. Explore, establish, exploit: Red teaming language models from scratch, 2023. URL: https://arxiv.org/abs/2306.09442, arXiv:2306.09442, doi:10.48550/arXiv.2306.09442.

[69] National Cyber Security Center. Introducing our new machine learning security principles, retrieved February 2023 from https://www.ncsc.gov.uk/blog-post/introducing-our-new-machine-learning-security-principles. URL: https://www.ncsc.gov.uk/blog-post/introducing-our-new-machine-learning-security-principles.

[70] Varun Chandrasekaran, Kamalika Chaudhuri, Irene Giacomelli, Somesh Jha, and Songbai Yan. Exploring connections between active learning and model extraction. In *Proceedings of the 29th USENIX Conference on Security Symposium*, SEC'20, USA, 2020. USENIX Association. URL: https://arxiv.org/abs/1811.02054, doi:10.48550/arXiv.1811.02054.

[71] Hong Chang, Ta Duy Nguyen, Sasi Kumar Murakonda, Ehsan Kazemi, and R. Shokri. On adversarial bias and the robustness of fair machine learning. https://arxiv.org/abs/2006.08669, 2020.

[72] Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J.

Sun 和 J.Zico Kolter。(已认证!!)免费对抗鲁棒性!,2023。URL:https://arxiv.org/abs/2206.10550, arXiv:2206.10550,doi:10.48550/arXiv.2 206.10550。

[63] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel.从大语言模型中提取训练数据。载于第 *30* 届 *USENIX* 安全研讨会*(USENIX Security 21)*，第 2633-2650 页。USENIX Association,2021 年 8 月。URL:https://www.usenix.org/conference/USENIX Security21/present ation/carlini-Extracting。

[64] Nicholas Carlini and David Wagner.对抗样本不易被检测：绕过十种检测方法。载于第 *10* 届 *ACM* 人工智能与安全研讨会论文集,AISec '17，第 3-14 页，New York, NY, USA,2017。Association for Computing Machinery。doi:10.1145/3128572.3140444.

[65] Nicholas Carlini and David Wagner.面向评估神经网络的鲁棒性。载于*Proc.IEEE Security and Privacy Symposium*,2017。URL:https://arxiv.org/abs/1608.04644,doi:10.48550/arXiv.1608.04644.

[66] Nicholas Carlini and David Wagner.音频对抗样本：对语音转文本的定向攻击。载于 *2018 IEEE Security and Privacy Workshops(SPW)*，第 1-7 页。IEEE,2018。URL:https://arxiv.org/abs/1801.01944, doi:10.48550/arXiv.1801. 01944。

[67] Stephen Casper, Yuxiao Li, Jiawei Li, Tong Bu, Kevin Zhang, Kaivalya Hariharan, and Dylan Hadfield-Menell.使用特征合成工具对深度神经网络进行红队测试。*arXiv* 预印本 *arXiv:2302.10894*,2023。URL:https://arxiv.org/abs/2302.108 94,doi:10.48550/arXiv.2302.10894.

[68] Stephen Casper, Jason Lin, Joe Kwon, Gatlen Culp，和 Dylan Hadfield-Menell。Explore, establish, exploit：从零开始对大语言模型进行红队测试，2023。URL: https://arxiv.org/abs/2306.09442, arXiv:2306.09442, doi:10.48550/arXiv.2 306.09442。

[69] 国家网络安全中心。介绍我们新的机器学习安全原则--原则，2023年2月检索自 https://www.ncsc.gov.uk/blog-post/introduc ing-our-new-machine-learning-security-principles。网址：https://www.ncsc.gov.u k/blog-post/introducing-our-new-machine-learning-security-principles。

[70] Varun Chandrasekaran, Kamalika Chaudhuri, Irene Giacomelli, Somesh Jha, and Songbai Yan.探索主动学习与模型提取之间的联系。载于第 *29* 届 *USENIX* 安全研讨会论文集,SEC'20,USA,2020。USENIX Association。URL:https://arxiv.org/abs/1811.02054, doi:10.48550/arXiv.1811.02054.

[71] Hong Chang, Ta Duy Nguyen, Sasi Kumar Murakonda, Ehsan Kazemi，和 R.Shokri. 论对抗性偏差与公平机器学习的鲁棒性.https://arxiv.org/ abs/2006.08669, 2020.

[72] 帕特里克·赵、爱德华多·德贝内代蒂、亚历山大·罗比、马克西姆·安德里乌申科、弗朗切斯科·克罗切、维卡什·塞瓦格、埃德加·多布里班、尼古拉·弗拉马里翁、乔治·J.

Pappas, Florian Tramer, Hamed Hassani, and Eric Wong. JailbreakBench: An open robustness benchmark for jailbreaking large language models, 2024. URL: https://arxiv.org/abs/2404.01318, arXiv:2404.01318, doi:10.48550/arXiv.2404.01318.

[73] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*, 2023. URL: https://arxiv.org/abs/2310.08419, doi:10.48550/arXiv.2310.08419.

[74] Harsh Chaudhari, John Abascal, Alina Oprea, Matthew Jagielski, Florian Tramèr, and Jonathan Ullman. SNAP: Efficient extraction of private properties with poisoning. In *2023 IEEE Symposium on Security and Privacy (S&P)*, 2023. URL: https://arxiv.org/abs/2208.12348, doi:10.48550/arXiv.2208.12348.

[75] Harsh Chaudhari, Giorgio Severi, John Abascal, Matthew Jagielski, Christopher A. Choquette-Choo, Milad Nasr, Cristina Nita-Rotaru, and Alina Oprea. Phantom: General trigger attacks on retrieval augmented language generation, 2024. URL: https://arxiv.org/abs/2405.20485, arXiv:2405.20485, doi:10.48550/arXiv.2405.20485.

[76] Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. Detecting backdoor attacks on deep neural networks by activation clustering. https://arxiv.org/abs/1811.03728, 2018. URL: https://arxiv.org/abs/1811.03728, doi:10.48550/arXiv.1811.03728.

[77] Hongge Chen, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, and Cho-Jui Hsieh. Attacking visual language grounding with adversarial examples: A case study on neural image captioning. https://arxiv.org/abs/1712.02051, 2017. URL: https://arxiv.org/abs/1712.02051, doi:10.48550/ARXIV.1712.02051.

[78] Huili Chen, Cheng Fu, Jishen Zhao, and Farinaz Koushanfar. DeepInspect: A black-box trojan detection and mitigation framework for deep neural networks. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 4658–4664. International Joint Conferences on Artificial Intelligence Organization, 7 2019. doi:10.24963/ijcai.2019/647.

[79] Jianbo Chen, Michael I. Jordan, and Martin J. Wainwright. HopSkipJumpAttack: A query-efficient decision-based attack. In *2020 IEEE Symposium on Security and Privacy, SP 2020, San Francisco, CA, USA, May 18-21, 2020*, pages 1277–1294. IEEE, 2020. doi:10.1109/SP40000.2020.00045.

[80] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, AISec '17, page 15–26, New York, NY, USA, 2017. Association for Computing Machinery. doi:10.1145/3128572.3140448.

[81] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Chau. *ShapeShifter: Robust Physical Adversarial Attack on Faster R-CNN Object Detector*, page 52–68. Springer International Publishing, 2019. URL: http://dx.doi.org/10.1007/978-3-030-10925-7_4, doi:10.1007/978-3-030-10925-7_4.

Pappas, Florian Tramer, Hamed Hassani, and Eric Wong.JailbreakBench： 一个用于越狱大型语言模型的开放鲁棒性基准， 2024. URL: https://arxiv.org/abs/2404.01318, arXiv:2404.01318, doi:10.48550/arXiv.2404.01318。

[73] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong.在二十次查询内越狱黑盒大语言模型。*arXiv*预印本 *arXiv:2310.08419*, 2023. URL: https://arxiv.org/abs/2310.08419, doi:10.48550/arXiv.2310.08419。

[74] Harsh Chaudhari, John Abascal, Alina Oprea, Matthew Jagielski, Florian Tramèr, and Jonathan Ullman.SNAP：通过投毒高效提取私有属性。载于*2023 IEEE*安全与隐私研讨会*(S&P)*, 2023. URL: https://arxiv.org/abs/2208.12348, doi:10.48550/arXiv.2208.12348。

[75] 严酷的乔杜里、乔治·塞韦里、约翰·阿巴斯卡、马修·贾吉尔斯基、克里斯托弗·A。Choquette-Choo、Milad Nasr、Cristina Nita-Rotaru 和 Alina Oprea。Phantom：对检索增强语言生成的通用触发攻击，2024。URL:https://arxiv.org/abs/2405.20485, arXiv:2405.20485,doi:10.48550/arXiv.2 405.20485。

[76] Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava.通过激活聚类检测深度神经网络上的后门攻击.https://arxiv.org/abs/1811.03728, 2018. URL: https://arxiv.org/abs/1811.03728, doi:10.48550/arXiv.1811.03728.

[77] Hongge Chen, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, and Cho-Jui Hsieh.用对抗样本攻击视觉语言接地： 神经图像描述案例研究.https://arxiv.org/abs/1712.02051, 2017. URL: https://arxiv.org/abs/17 12.02051, doi:10.48550/ARXIV.1712.02051.

[78] Huili Chen, Cheng Fu, Jishen Zhao, and Farinaz Koushanfar.DeepInspect：一种用于深度神经网络的黑盒木马检测与缓解框架。载于第二十八届国际人工智能联合会议论文集，*IJCAI-19*，第4658-4664页。国际人工智能联合会议组织，2019年7月。doi:10.24963/ijcai.2019/647.

[79] Jianbo Chen, Michael I.Jordan, and Martin J.Wainwright.HopSkipJumpAttack：一种查询高效的基于决策的攻击。载于*2020 IEEE*安全与隐私研讨会，*SP 2020*，美国加利福尼亚州旧金山，*2020*年*5*月*18-21*日，第1277-1294页。IEEE,2020。doi:10.1109/SP40000.2020.00045.

[80] 陈品宇，张欢，亚什·夏尔马，易金凤，谢卓睿。动物园：第零或-基于优化的黑盒攻击深度神经网络而无需训练替代模型。载于第*10*届*ACM*人工智能与安全研讨会论文集,AISec '17，第15-26页，美国纽约州纽约市，2017年。美国计算机协会。doi:10.1145/3128572.3140448.

[81] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Chau.*ShapeShifter*：针对*Faster R-CNN*目标检测器的鲁棒物理对抗攻击，第52-68页。Springer International Publishing,2019年。URL: http://dx.doi.org/10.1007/978-3-030 -10925-7_4, doi:10.1007/978-3-030-10925-7_4.

[82] Xiaoyi Chen, Ahmed Salem, Dingfan Chen, Michael Backes, Shiqing Ma, Qingni Shen, Zhonghai Wu, and Yang Zhang. Badnl: Backdoor attacks against nlp models with semantic-preserving improvements. In *Annual Computer Security Applications Conference*, ACSAC '21, page 554–569, New York, NY, USA, 2021. Association for Computing Machinery. doi:10.1145/3485832.3485837.

[83] Xiaoyi Chen, Siyuan Tang, Rui Zhu, Shijun Yan, Lei Jin, Zihao Wang, Liya Su, Zhikun Zhang, XiaoFeng Wang, and Haixu Tang. The Janus interface: How fine-tuning in large language models amplifies the privacy risks, 2024. URL: https://arxiv.org/abs/2310.15469, arXiv:2310.15469, doi:10.48550/arXiv.2310.15469.

[84] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017. URL: https://arxiv.org/abs/1712.05526, doi:10.48550/arXiv.1712.05526.

[85] Heng-Tze Cheng and Romal Thoppilan. LaMDA: Towards Safe, Grounded, and High-Quality Dialog Models for Everything. https://ai.googleblog.com/2022/01/lamda-towards-safe-grounded-and-high.html, 2022. Google Brain. URL: https://research.google/blog/lamda-towards-safe-grounded-and-high-quality-dialog-models-for-everything/.

[86] Minhao Cheng, Thong Le, Pin-Yu Chen, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. Query-efficient hard-label black-box attack: An optimization-based approach. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL: https://openreview.net/forum?id=rJlk6iRqKX, doi:10.48550/arXiv.1807.04457.

[87] Minhao Cheng, Simranjit Singh, Patrick H. Chen, Pin-Yu Chen, Sijia Liu, and Cho-Jui Hsieh. Sign-opt: A query-efficient hard-label adversarial attack. In *International Conference on Learning Representations*, 2020. URL: https://openreview.net/forum?id=SklTQCNtvS, doi:10.48550/arXiv.1909.10773.

[88] Alesia Chernikova and Alina Oprea. FENCE: Feasible evasion attacks on neural networks in constrained environments. *ACM Transactions on Privacy and Security (TOPS) Journal*, 2022. URL: https://arxiv.org/abs/1909.10480, doi:10.48550/arXiv.1909.10480.

[89] Christopher A. Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. Label-only membership inference attacks. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1964–1974. PMLR, 18–24 Jul 2021. URL: https://proceedings.mlr.press/v139/choquette-choo21a.html, doi:10.48550/arXiv.2007.14321.

[90] Sheng-Yen Chou, Pin-Yu Chen, and Tsung-Yi Ho. How to backdoor diffusion models? https://arxiv.org/abs/2212.05400, 2022. doi:10.48550/ARXIV.2212.05400.

[91] Antonio Emanuele Cinà, Kathrin Grosse, Ambra Demontis, Sebastiano Vascon, Werner Zellinger, Bernhard A. Moser, Alina Oprea, Battista Biggio, Marcello Pelillo, and Fabio Roli. Wild patterns reloaded: A survey of machine learning security

[82] Xiaoyi Chen、Ahmed Salem、Dingfan Chen、Michael Backes、Shiqing Ma、Qingni Shen、Zhonghai Wu 和 Yang Zhang。Badnl：针对 NLP 模型的、具有语义保持改进的后门攻击。载于*Annual Computer Security Applications Conference*,ACSAC '21，第 554-569 页，美国纽约州纽约市，2021 年。Association for Computing Machinery。doi:10.1145/3485832.3485837.

[83] Xiaoyi Chen, Siyuan Tang, Rui Zhu, Shijun Yan, Lei Jin, Zihao Wang, Liya Su, Zhikun Zhang, XiaoFeng Wang, and Haixu Tang.Janus接口： 大型语言模型中的微调如何放大隐私风险， 2024. URL: https://arxiv.org/ab s/2310.15469, arXiv:2310.15469, doi:10.48550/arXiv.2310.15469.

[84] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song.使用数据投毒对深度学习系统的定向后门攻击。*arXiv*预印本 *arXiv:1712.05526*, 2017. URL: https://arxiv.org/abs/1712.05526, doi:10.48550/arXiv.1712.05526.

[85] Heng-Tze Cheng and Romal Thoppilan.LaMDA： 面向安全、有据且高质量的对话模型.https://ai.googleblog.com/2022/01/la mda-towards-safe-grounded-and-high.html, 2022. Google Brain.URL: https: //research.google/blog/lamda-towards-safe-grounded-and-high-quality-dialog-m odels-for-Everything/.

[86] Minhao Cheng, Thong Le, Pin-Yu Chen, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh.查询高效的硬标签黑盒攻击：一种基于优化的方法。载于第*7*届国际学习表征会议，*ICLR 2019*，美国路易斯安那州新奥尔良，*2019*年*5*月*6-9*日。OpenReview.net,2019年。URL: https://OpenReview.net/forum?i

d=rJlk6iRqKX，doi:10.48550/arXiv.1807.04457.

[87] Minhao Cheng、Simranjit Singh、Patrick H.Chen、Pin-Yu Chen、Sijia Liu 和 Cho-Jui Hsieh。Sign-opt：一种查询高效的硬标签对抗攻击。载于《国际学习表征会议》，2020年。网址：https://openreview.net/for umid=SklTQCNtvS,doi:10.48550/arXiv.1909.10773.

[88] Alesia Chernikova and Alina Oprea.FENCE： 神经网络上的可行规避攻击在受限环境中的工作。*ACM*隐私与安全汇刊*(TOPS)*期刊, 2022. URL: https://arxiv.org/abs/1909.10480, doi:10.48550/arXiv.1909.10480.

[89] Christopher A.Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot.仅标签的成员推断攻击。载于Marina Meila和Tong Zhang编辑，第*38*届国际机器学习会议论文集，第139卷，机器学习研究论文集，第1964-1974页。PMLR,2021年7月18-24日。URL: https://Proceedings.mlr.press/v139/choquette-choo21a.html,

doi:10.48550/arXiv.2007.14321。

[90] 周圣彦、陈品谕与何宗易。如何对扩散模型进行后门攻击?
https://arxiv.org/abs/2212.05400，2022年。doi:10.48550/ARXIV.2212.05400。

[91] Antonio Emanuele Cinà, Kathrin Grosse, Ambra Demontis, Sebastiano Vascon, Werner Zellinger, Bernhard A.Moser, Alina Oprea, Battista Biggio, Marcello Pelillo, and Fabio Roli.野性模式重载： 机器学习安全

against training data poisoning. *ACM Computing Surveys*, March 2023. URL: https://doi.org/10.1145%2F3585385, doi:10.1145/3585385.

[92] Jack Clark and Raymond Perrault. 2022 AI index report. https://aiindex.stanford.edu/wp-content/uploads/2022/03/2022-AI-Index-Report_Master.pdf, 2022. Human Centered AI, Stanford University.

[93] Joseph Clements, Yuzhe Yang, Ankur Sharma, Hongxin Hu, and Yingjie Lao. Rallying adversarial techniques against deep learning for network security, 2019. URL: https://arxiv.org/abs/1903.11688, doi:10.48550/ARXIV.1903.11688.

[94] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1310–1320. PMLR, 09–15 Jun 2019. URL: https://proceedings.mlr.press/v97/cohen19c.html.

[95] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pages 1310–1320. PMLR, 2019.

[96] Gabriela F. Cretu, Angelos Stavrou, Michael E. Locasto, Salvatore J. Stolfo, and Angelos D. Keromytis. Casting out demons: Sanitizing training data for anomaly sensors. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 81–95, 2008. URL: https://ieeexplore.ieee.org/document/4531146, doi:10.1109/SP.2008.11.

[97] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL: https://openreview.net/forum?id=SSKZPJCt7B, doi:10.48550/arXiv.2010.09670.

[98] Nilesh Dalvi, Pedro Domingos, Mausam, Sumit Sanghai, and Deepak Verma. Adversarial classification. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 99–108, New York, NY, USA, 2004. Association for Computing Machinery. doi:10.1145/1014052.1014066.

[99] DARPA. DARPA AI Cyber Challenge Aims to Secure Nation's Most Critical Software, 2023. Accessed: 2024-08-22. URL: https://www.darpa.mil/news-events/2023-08-09.

[100] Emiliano De Cristofaro. A critical overview of privacy in machine learning. *IEEE Security & Privacy*, 19(4):19–27, 2021. doi:10.1109/MSEC.2021.3076443.

[101] Edoardo Debenedetti, Jie Zhang, Mislav Balunović, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. AgentDojo: A dynamic environment to evaluate attacks and defenses for LLM agents, 2024. URL: https://arxiv.org/abs/2406.13352, arXiv:2406.13352, doi:10.48550/arXiv.2406.13352.

[102] DeepMind. Building safer dialogue agents. https://www.deepmind.com/blog/building-safer-dialogue-agents, 2022. Online.

针对训练数据投毒。*ACM*计算综述,2023年3月。URL: https://doi.org/10.1145%2F3585385, doi:10.1145/3585385.

[92] 杰克·克拉克和雷蒙德·佩罗。2022年AI指数报告。https://aiindex.stanford.edu/wp-content/uploads/2022/03/2022-AI-Index-Report_Master.pdf,2022年。斯坦福大学以人为本人工智能研究院。

[93] Joseph Clements, Yuzhe Yang, Ankur Sharma, Hongxin Hu, and Yingjie Lao.集结对抗技术对抗深度学习用于网络安全， 2019. URL: https: //arxiv.org/abs/1903.11688, doi:10.48550/ARXIV.1903.11688.

[94] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter.通过随机平滑实现可认证的对抗鲁棒性。载于Kamalika Chaudhuri和Ruslan Salakhutdinov编辑，第*36*届国际机器学习会议论文集，第97卷，机器学习研究论文集，第1310-1320页。PMLR,6月09-15日

2019. 网址：https://proceedings.mlr.press/v97/cohen19c.html.

[95] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter.通过随机平滑实现可认证的对抗鲁棒性。载于国际机器学习会议，第1310-1320页。PMLR,2019年。

[96] 加布里埃拉·F。克雷图，安杰洛斯·斯塔夫鲁，迈克尔·E。洛卡斯托，萨尔瓦托雷·J。斯托尔福，还有安格-

los D.Keromytis.驱除恶魔：为异常传感器净化训练数据。载于*2008 IEEE*安全与隐私研讨会*(sp 2008)*，第81-95页，2008年。URL: https://ieeexplore.ieee.org/document/4531146, doi:10.1109/SP.2008.11.

[97] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein.Robustbench：一个标准化的对抗鲁棒性基准。载于第三十五届神经信息处理系统会议数据集与基准赛道*(*第*2*轮*)*, 2021. URL: https://openreview.net/forum?id=SSKZPJCt7B, doi: 10.48550/arXiv.2010.09670.

[98] Nilesh Dalvi、Pedro Domingos、Mausam、Sumit Sanghai 和 Deepak Verma。Adver-

对抗分类。载于第十届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集,KDD '04，第99-108页，美国纽约州纽约市，2004年。美国计算机协会。doi:10.1145/1014052.10

[99] DARPA。DARPA AI网络挑战赛旨在保障国家最关键软件的安全，2023。访问日期：2024-08-22。URL:https://www.darpa.mil/news-events/2023-0 8-09。

[100] Emiliano De Cristofaro.机器学习中隐私的批判性综述。*IEEE*安全与隐私,19(4): 19-27, 2021. doi:10.1109/MSEC.2021.3076443.

[101] 爱德华多·德贝内代蒂，张杰，米斯拉夫·巴卢诺维奇，卢卡·博伊雷尔·凯尔纳，马克·菲斯-

和 Florian Tramèr。AgentDojo：用于评估LLM智能体攻击与防御的动态环境，2024。URL:https://arxiv.org/abs/2406.13352, arXiv:2406.13352,doi:10.48550/arXiv.2406.13352.

[102] DeepMind。构建更安全的对话智能体。https://www.deepmind.com/blog/building-safer-dialogue-agents,2022年。在线。

[103] Luca Demetrio, Battista Biggio, Giovanni Lagorio, Fabio Roli, and Alessandro Armando. Functionality-preserving black-box optimization of adversarial windows malware. *IEEE Transactions on Information Forensics and Security*, 16:3469–3478, 2021. URL: https://ieeexplore.ieee.org/document/9437194, doi:10.1109/TIFS.2021.3082330.

[104] Ambra Demontis, Marco Melis, Maura Pintor, Matthew Jagielski, Battista Biggio, Alina Oprea, Cristina Nita-Rotaru, and Fabio Roli. Why do adversarial attacks transfer? Explaining transferability of evasion and poisoning attacks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 321–338. USENIX Association, 2019. URL: https://www.usenix.org/conference/usenixsecurity19/presentation/demontis.

[105] Serguei Denissov, Hugh Brendan McMahan, J Keith Rush, Adam Smith, and Abhradeep Guha Thakurta. Improved differential privacy for SGD via optimal private linear operators on adaptive streams. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL: https://openreview.net/forum?id=i9XrHJoyLqJ, doi:10.48550/arXiv.2202.08312.

[106] Leon Derczynski. Garak: LLM vulnerability scanner. https://github.com/leondz/garak, 2024. Accessed: 2024-08-18.

[107] Leon Derczynski, Erick Galinkin, Jeffrey Martin, Subho Majumdar, and Nanna Inie. garak: A framework for security probing large language models, 2024. URL: https://arxiv.org/abs/2406.11036, arXiv:2406.11036, doi:10.48550/arXiv.2406.11036.

[108] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL: https://openreview.net/forum?id=OUIFPHEgJU, doi:10.48550/arXiv.2305.14314.

[109] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. In *International Conference on Machine Learning*, pages 1596–1606. PMLR, 2019. URL: https://arxiv.org/abs/1803.02815, doi:10.48550/arXiv.1803.02815.

[110] Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the 22nd ACM Symposium on Principles of Database Systems*, PODS '03, pages 202–210. ACM, 2003. URL: https://crypto.stanford.edu/seclab/sem-03-04/psd.pdf.

[111] Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush R. Varshney. Is there a trade-off between fairness and accuracy? A perspective using mismatched hypothesis testing. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020. URL: https://arxiv.org/abs/1910.07870, doi:10.48550/arXiv.1910.07870.

[112] Cynthia Dwork. Differential privacy. In *Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceed-*

[103] 卢卡·德梅特里奥、巴蒂斯塔·比吉奥、乔瓦尼·拉戈里奥、法比奥·罗利和亚历山德罗·阿尔-mando.对抗性Windows恶意软件的功能保持黑盒优化。*IEEE*信息取证与安全汇刊,16:3469-3478, 2021. URL: https://ieeexplore.ieee.org/document/9437194, doi:10.1109/TIFS.2021.3082330.

[104] Ambra Demontis, Marco Melis, Maura Pintor, Matthew Jagielski, Battista Biggio, Alina Oprea, Cristina Nita-Rotaru, and Fabio Roli.对抗攻击为何会迁移?解释规避和投毒攻击的可迁移性。载于第*28*届*USENIX*安全研讨会*(USENIX Security 19)*，第321-338页。USENIX协会，2019年。URL: https://www.usenix.org/conference/USENIX Security19/presentation/demont

是。

[105] Serguei Denissov, Hugh Brendan McMahan, J Keith Rush, Adam Smith, and Abhradeep Guha Thakurta.通过自适应流上的最优私有线性算子改进SGD的差分隐私。载于Alice H.Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho编辑，神经信息处理系统进展, 2022. URL: https://openreview.net/forum?id=i9XrHJoyLqJ,

doi:10.48550/arXiv.2202.08312。

[106] Leon Derczynski。Garak:LLM 漏洞扫描器。https://github.com/leondz/garak,2024。访问日期：2024-08-18。

[107] Leon Derczynski, Erick Galinkin, Jeffrey Martin, Subho Majumdar, and Nanna Inie。garak：用于大语言模型安全探测的框架，2024。URL:https://arxiv.org/abs/2406.11036,arXiv:2406.11036,doi:10.48550/arXiv.2406. 11036。

[108] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer。QLoRA：量化LLM的高效微调。

量化LLM的高效微调。载于第三十七届神经信息处理系统会议, 2023. URL: https://openreview.net/forum?id=OUIFPH

[109] 伊利亚斯·迪亚科尼科拉斯、高塔姆·卡马斯、丹尼尔·凯恩、杰瑞·李、雅各布·斯坦哈特和阿尔-istair Stewart.Sever：一种用于随机优化的鲁棒元算法。载于国际机器学习会议，第1596-1606页。PMLR,2019年。URL: https://arxiv.org/abs/1803.02815, doi:10.48550/arXiv.1803.02815.

[110] Irit Dinur and Kobbi Nissim。在保护隐私的同时揭示信息。载于*Proceedings of the 22nd ACM Symposium on Principles of Database*

第*22*届*ACM*数据库系统原理研讨会论文集,PODS '03，第202-210页。ACM,2003年。URL: https://crypto.stanford.edu/seclab/sem-03-04/psd.pdf.

[111] Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush R.Varshney.公平性与准确性之间是否存在权衡?基于失配假设检验的视角。载于第*37*届国际机器学习会议论文集,ICML'20。JMLR.org,2020年。URL: https://arxiv.org/abs/1910.0 7870, doi:10.48550/arXiv.1910.07870.

[112] Cynthia Dwork.差分隐私。载于自动机、语言与编程，第*33*届国际学术讨论会，*ICALP 2006*，意大利威尼斯，*2006*年*7*月*10-14*日，论文-

ings, Part II, pages 1–12, 2006. URL: http://dx.doi.org/10.1007/11787006_1, doi:10.1007/11787006_1.

[113] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Conference on Theory of Cryptography*, TCC '06, pages 265–284, New York, NY, USA, 2006.

[114] Cynthia Dwork, Adam Smith, Thomas Steinke, and Jonathan Ullman. Exposed! A survey of attacks on private data. *Annual Review of Statistics and Its Application*, 4:61–84, 2017. URL: https://privacytools.seas.harvard.edu/publications/exposed-survey-attacks-private-data.

[115] Cynthia Dwork, Adam Smith, Thomas Steinke, Jonathan Ullman, and Salil Vadhan. Robust traceability from trace amounts. In *IEEE Symposium on Foundations of Computer Science*, FOCS '15, 2015. URL: https://privacytools.seas.harvard.edu/files/privacytools/files/robust.pdf.

[116] Cynthia Dwork and Sergey Yekhanin. New efficient attacks on statistical disclosure control mechanisms. In *Annual International Cryptology Conference*, pages 469–480. Springer, 2008. URL: https://link.springer.com/chapter/10.1007/978-3-540-85174-5_26, doi:0.1007/978-3-540-85174-5_26.

[117] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. Hotflip: White-box adversarial examples for text classification. *arXiv preprint arXiv:1712.06751*, 2017. URL: https://arxiv.org/abs/1712.06751, doi:10.48550/arXiv.1712.06751.

[118] Kazuki Egashira, Mark Vero, Robin Staab, Jingxuan He, and Martin Vechev. Exploiting LLM Quantization, 2024. URL: https://arxiv.org/abs/2405.18137, arXiv:2405.18137, doi:10.48550/arXiv.2405.18137.

[119] Gemini Team et al. Gemini: A family of highly capable multimodal models. https://arxiv.org/abs/2312.11805, 2023. arXiv:2312.11805.

[120] ETSI Group Report SAI 005. Securing artificial intelligence (SAI); mitigation strategy report, retrieved February 2023 from https://www.etsi.org/deliver/etsi_gr/SAI/001_099/005/01.01.01_60/gr_SAI005v010101p.pdf. URL: https://www.etsi.org/deliver/etsi_gr/SAI/001_099/005/01.01.01_60/gr_SAI005v010101p.pdf.

[121] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1625–1634, 2018. URL: https://ieeexplore.ieee.org/document/8578273, doi:10.1109/CVPR.2018.00175.

[122] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1625–1634. Computer Vision Foundation / IEEE Computer Society, 2018. URL: http://openaccess.thecvf.com/content_cvpr_2018/html/Eykholt_Robust_Physical-World_Attacks_CVPR_2018_paper.html, doi:10.1109/CVPR.2018.00175.

[123] Minghong Fang, Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. Local Model Poi-

ings，第二部分，第1-12页，2006年。网址：http://dx.doi.org/10.1007/11787006_1, doi:10.1007/11787006_1。

[113] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith.将噪声校准至私有数据分析中的敏感度。载于密码学理论会议,TCC '06，第265-284页，美国纽约州纽约市，2006年。

[114] Cynthia Dwork、Adam Smith、Thomas Steinke 和 Jonathan Ullman。曝光！隐私数据攻击综述。《统计及其应用年度评论》，4:61-84，2017。网址：https://privacytools.seas.harvard.edu/publications/exposed -survey-attacks-private-data。

[115] Cynthia Dwork、Adam Smith、Thomas Steinke、Jonathan Ullman 和 Salil Vadhan。从微量数据中实现稳健的可追溯性。载于*IEEE* 计算机科学基础研讨会,FOCS '15,2015 年。URL:https://privacytools.seas.harvard.edu/files/pr ivacytools/files/robust.pdf。

[116] Cynthia Dwork and Sergey Yekhanin。对统计披露控制机制的新型高效攻击。控制机制。载于年度国际密码学会议，第469-480页。Springer,2008年。网址：https://link.springer.com/chapter/10.1007/978-3-5

[117] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou。Hotflip：用于文本分类的白盒对抗样本。用于文本分类的对抗样本。*arXiv*预印本 *arXiv:1712.06751*,2017年。网址：https://arxiv.org/abs/1712.06751, doi:10.48550/arXiv.1712.06751.

[118] [75] Kazuki Egashira、Mark Vero、Robin Staab、Jingxuan He 和 Martin Vechev。利用LLM量化，2024年。网址：(https://arxiv.org/abs/2405.18137,) arXiv:2405.1 8137，doi:(10.48550/arXiv.2405.18137.)

[119] Gemini Team等。Gemini：一系列能力强大的多模态模型。https://arxiv.org/abs/2312.11805，2023年。arXiv:2312.11805。

[120] ETSI Group Report SAI 005。保障人工智能(SAI)安全；缓解策略报告，2023年2月取自https://www.etsi.org/deliver/etsi_gr/SAI/0 01_099/005/01.01.01_60/gr_SAI005v010101p.pdf。URL:https://www.org/del iver/etsi_gr/SAI/001_099/005/01.01.01_60/gr_SAI005v010101p.pdf。

[121] Kevin Eykholt、Ivan Evtimov、Earlence Fernandes、Bo Li、Amir Rahmati、Chaowei Xiao、Atul Prakash、Tadayoshi Kohno 和 Dawn Song。针对深度学习视觉分类的稳健物理世界攻击。载于*2018 IEEE/CVF* 计算机视觉与模式识别会议，第 1625-1634 页，2018 年。网址：https://ieeexplore.ieee.or g/document/8578273,doi:10.1109/CVPR.2018.00175.

[122] Kevin Eykholt、Ivan Evtimov、Earlence Fernandes、Bo Li、Amir Rahmati、Chaowei Xiao、Atul Prakash、Tadayoshi Kohno 和 Dawn Song。针对深度学习视觉分类的鲁棒物理世界攻击。载于*2018 IEEE* 计算机视觉与模式识别会议，*CVPR 2018*，美国犹他州盐湖城，*2018*年*6*月*18-22*日，第1625-1634页。计算机视觉基金会 / IEEE 计算机学会，2018年。URL: http: //openaccess.thecvf.com/content_cvpr_2018/html/Eykholt_Robust_Physical-Wor ld_Attacks_CVPR_2018_paper.html, doi:10.1109/CVPR.2018.00175.

[123] Minghong Fang、Xiaoyu Cao、Jinyuan Jia 和 Neil Zhenqiang Gong。本地模型投毒

soning Attacks to Byzantine-Robust Federated Learning. In *USENIX Security*, 2020.

[124] Zhen Fang, Yixuan Li, Jie Lu, Jiahua Dong, Bo Han, and Feng Liu. Is out-of-distribution detection learnable? In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022)*. online: https://arxiv.org/abs/2210.14707, 2022. doi:10.48550/ARXIV.2210.14707.

[125] Georgios Fatouros, John Soldatos, Kalliopi Kouroumali, Georgios Makridis, and Dimosthenis Kyriazis. Transforming sentiment analysis in the financial domain with chatgpt. *Machine Learning with Applications*, 14:100508, 2023. URL: https://www.sciencedirect.com/science/article/pii/S2666827023000610, doi:10.1016/j.mlwa.2023.100508.

[126] Vitaly Feldman. Does learning require memorization? A short tale about a long tail. In *ACM Symposium on Theory of Computing*, STOC '20, pages 954–959, 2020. https://arxiv.org/abs/1906.05271.

[127] Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. URL: https://arxiv.org/abs/2008.03703, doi:10.48550/arXiv.2008.03703.

[128] Ji Feng, Qi-Zhi Cai, and Zhi-Hua Zhou. Learning to confuse: Generating training time adversarial data with auto-encoder. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/file/1ce83e5d4135b07c0b82afffbe2b3436-Paper.pdf, doi:10.48550/arXiv.1905.09027.

[129] Liam Fowl, Ping-yeh Chiang, Micah Goldblum, Jonas Geiping, Arpit Bansal, Wojtek Czaja, and Tom Goldstein. Preventing unauthorized use of proprietary data: Poisoning for secure dataset release, 2021. URL: https://arxiv.org/abs/2103.02683, doi:10.48550/ARXIV.2103.02683.

[130] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, CCS '15, page 1322–1333, New York, NY, USA, 2015. Association for Computing Machinery. doi:10.1145/2810103.2813677.

[131] Aymeric Fromherz, Klas Leino, Matt Fredrikson, Bryan Parno, and Corina Pasareanu. Fast geometric projections for local robustness certification. In *International Conference on Learning Representations*, 2021. URL: https://openreview.net/forum?id=zWy1uxjDdZJ, doi:10.48550/arXiv.2002.04742.

[132] Pranav Gade, Simon Lermen, Charlie Rogers-Smith, and Jeffrey Ladish. Badllama: cheaply removing safety fine-tuning from llama 2-chat 13b, 2024. URL: https://arxiv.org/abs/2311.00117, arXiv:2311.00117, doi:10.48550/arXiv.2311.00117.

[133] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones,

攻击.In *USENIX Security*, 2020.

[124] 方振、李奕萱、陆杰、董佳华、韩波和刘峰。分布外检测是可学习的吗?见第*36*届神经信息处理系统大会*(NeurIPS 2022)*论文集。在线：https://arxiv.org/abs/2210.14707, 2022。doi:10.48550/ARXIV.2210.14707.

[125] Georgios Fatouros, John Soldatos, Kalliopi Kouroumali, Georgios Makridis，和 Dimosthenis Kyriazis。使用 chatgpt 转变金融领域的情感分析。*Machine Learning with Applications*,14:100508,2023。URL:https://www.sciencedirect.com/science/article/pii/S2666827023000610,doi:

10.1016 /j.

[126] Vitaly Feldman。学习需要记忆吗?一个关于长尾的短故事。载于*ACM Symposium on Theory of Computing*,STOC '20，第954-959页，2020年。https://arxiv.org/abs/1906.05271.

[127] Vitaly Feldman and Chiyuan Zhang.神经网络记忆什么以及为什么：Dis-通过影响估计覆盖长尾。载于第*34*届神经信息处理系统国际会议论文集,NIPS'20，美国纽约州红钩，2020年。Curran Associates Inc.网址：https://arxiv.org/abs/2008.03703,

[128] Ji Feng、Qi-Zhi Cai 和 Zhi-Hua Zhou。学习混淆：用自编码器生成训练时对抗数据。载于 H.Wallach、H.Larochelle、A.Beygelzimer、F.d'Alché-Buc、E.Fox 和 R.Garnett 编，*Advances in Neural Information Processing Systems*，第 32 卷。Curran Associates, Inc.,2019 年。URL:https://proceedi ngs.neurips.cc/paper/2019/file/1ce83e5d4135b07c0b82afffbe2b3436-Paper.pdf,

doi:10.48550/arXiv.1905.09027。

[129] Liam Fowl, Ping-yeh Chiang, Micah Goldblum, Jonas Geiping, Arpit Bansal, Wojtek Czaja, and Tom Goldstein.防止专有数据的未经授权使用：用于安全数据集发布的投毒，2021。URL: https://arxiv.org/abs/2103.02683,

doi:10.48550/ARXIV.2103.02683。

[130] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart.利用置信度信息的模型反演攻击及基本对策。载于第*22*届*ACM SIGSAC*计算机与通信安全会议论文集,CCS '15，第1322-1333页，美国纽约州纽约市，2015年。美国计算机协会。

doi:10.1145/2810103.2813677。

[131] Aymeric Fromherz、Klas Leino、Matt Fredrikson、Bryan Parno 和 Corina Pasareanu。用于局部鲁棒性认证的快速几何投影。载于国际学习表征会议,2021年。网址：https://openreview.net/forum?id=zWy1uxjDdZJ，doi:10.48550/arXiv.2002.04742.

[132] Pranav Gade、Simon Lermen、Charlie Rogers-Smith 和 Jeffrey Ladish。Badllama：低成本移除 llama 2-chat 13b 的安全微调，2024。URL:https://arxi v.org/abs/2311.00117,arXiv:2311.00117,doi:10.48550/arXiv.2311.00117.

[133] 迪普·甘古利、莉安·洛维特、杰克逊·克尼恩、阿曼达·阿斯凯尔、白元涛、索拉夫·卡-达瓦思、本·曼、伊桑·佩雷斯、尼古拉斯·希弗、卡迈勒·恩杜斯、安迪·琼斯，

Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned, 2022. URL: https://arxiv.org/abs/2209.07858, arXiv:2209.07858, doi:10.48550/arXiv.2209.07858.

[134] Karan Ganju, Qi Wang, Wei Yang, Carl A. Gunter, and Nikita Borisov. Property inference attacks on fully connected neural networks using permutation invariant representations. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, CCS '18, page 619–633, New York, NY, USA, 2018. Association for Computing Machinery. doi:10.1145/3243734.3243834.

[135] Simson Garfinkel, John Abowd, and Christian Martindale. Understanding database reconstruction attacks on public data. *Communications of the ACM*, 62:46–53, 02 2019. URL: https://dl.acm.org/doi/10.1145/3287287, doi:10.1145/3287287.

[136] Timon Gehr, Matthew Mirman, Dana Drachsler-Cohen, Petar Tsankov, Swarat Chaudhuri, and Martin Vechev. AI2: Safety and robustness certification of neural networks with abstract interpretation. In *2018 IEEE Symposium on Security and Privacy (S&P)*, pages 3–18, 2018. URL: https://ieeexplore.ieee.org/document/8418593, doi:10.1109/SP.2018.00058.

[137] Jonas Geiping, Liam H Fowl, W. Ronny Huang, Wojciech Czaja, Gavin Taylor, Michael Moeller, and Tom Goldstein. Witches' brew: Industrial scale data poisoning via gradient matching. In *International Conference on Learning Representations*, 2021. URL: https://openreview.net/forum?id=01olnfLIbD, doi:10.48550/arXiv.2009.02276.

[138] Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W. Mahoney, and Kurt Keutzer. A survey of quantization methods for efficient neural network inference, 2021. URL: https://arxiv.org/abs/2103.13630, arXiv:2103.13630, doi:10.48550/arXiv.2103.13630.

[139] Antonio A. Ginart, Melody Y. Guan, Gregory Valiant, and James Zou. Making ai forget you: Data deletion in machine learning. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019. Curran Associates Inc. URL: https://arxiv.org/abs/1907.05012, doi:10.48550/arXiv.1907.05012.

[140] David Glukhov, Ilia Shumailov, Yarin Gal, Nicolas Papernot, and Vardan Papyan. LLM censorship: A machine learning challenge or a computer security problem?, 2023. URL: https://arxiv.org/abs/2307.10719, arXiv:2307.10719, doi:10.48550/arXiv.2307.10719.

[141] Micah Goldblum, Avi Schwarzschild, Ankit Patel, and Tom Goldstein. Adversarial attacks on machine learning systems for high-frequency trading. In *Proceedings of the Second ACM International Conference on AI in Finance*, ICAIF '21, New York, NY,

Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark.红队测试语言模型以减少危害：方法、扩展行为和经验教训， 2022. URL: https://arxiv.org/abs/2209.07858, arXiv:2209.07858,

doi:10.48550/arXiv.2209.07858。

[134] Karan Ganju、Qi Wang、Wei Yang、Carl A.Gunter 和 Nikita Borisov。属性推断-使用置换不变表示对全连接神经网络进行成员推理攻击。载于*2018*年*ACM SIGSAC*计算机与通信安全会议论文集,CCS '18，第619-633页，美国纽约州纽约市，2018年。美国计算机协会。doi:10.1145/3243734.3243834.

[135] https://arxi Simson Garfinkel、John Abowd 和 Christian Martindale。理解针对公共数据的数据库重构攻击。*Communications of the ACM*,62:46-53,2019年2月。URL:https://dl.acm.org/doi/10.1145/3287287, doi:10.1145/3287287.

[136] Timon Gehr、Matthew Mirman、Dana Drachsler-Cohen、Petar Tsankov、Swarat Chaudhuri 和 Martin Vechev。AI2：基于抽象解释的神经网络安全性与鲁棒性认证。载于*2018 IEEE Symposium on Security and Pri- vacy (S&P)*，第 3-18 页，2018 年。URL:https://ieeexplore.ieee.org/document/84185 93,doi:10.1109/SP.2018.00058.

[137] Jonas Geiping、Liam H Fowl、W.Ronny Huang、Wojciech Czaja、Gavin Taylor、Michael Moeller 和 Tom Goldstein。Witches' brew：通过梯度匹配实现的工业规模数据投毒(poisoning)。载于国际学习表征会议,2021 年。URL:https://openreview.net/forum?id=01olnfLIbD, doi:10.48550/arXiv.2 009.02276。

[138] Amir Gholami、Sehoon Kim、Zhen Dong、Zhewei Yao、Michael W.Mahoney 和 Kurt Keutzer。面向高效神经网络推理的量化方法综述，2021年。网址：https://arxiv.org/abs/2103.13630, arXiv:2103.13630,doi:10.485 50/arXiv.2103.13630。

[139] Antonio A.Ginart、Melody Y.Guan、Gregory Valiant 和 James Zou。让 AI 忘记你：机器学习中的数据删除。载于第 *33* 届国际神经信息处理系统大会论文集，美国纽约州红钩，2019 年。Curran Associates Inc.URL:https://arxiv.org/abs/1907.05012, doi:10.48550/a rXiv.1907.05012。

[140] David Glukhov、Ilia Shumailov、Yarin Gal、Nicolas Papernot 和 Vardan Papyan。LLM 审查：机器学习挑战还是计算机安全问题?,2023。URL:https://arxiv.org/abs/2307.10719, arXiv:2307.10719,doi:10.48550/arX iv.2307.10719。

[141] Micah Goldblum, Avi Schwarzschild, Ankit Patel, and Tom Goldstein.针对高频交易机器学习系统的对抗攻击。载于第二届*ACM*金融人工智能国际会议论文集,ICAIF '21，美国纽约州纽约市，

USA, 2021. Association for Computing Machinery. doi:10.1145/3490354.3494367.

[142] Shafi Goldwasser, Michael P. Kim, Vinod Vaikuntanathan, and Or Zamir. Planting undetectable backdoors in machine learning models. https://arxiv.org/abs/2204.06974, 2022. arXiv. doi:10.48550/ARXIV.2204.06974.

[143] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[144] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. URL: http://arxiv.org/abs/1412.6572, doi:10.48550/arXiv.1412.6572.

[145] Kai Greshake. Prompt injection defenses should suck less. https://kai-greshake.de/posts/approaches-to-pi-defense/, March 2024. Accessed: 2024-08-22.

[146] Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. *arXiv preprint arXiv:2302.12173*, 2023. URL: https://arxiv.org/abs/2302.12173, doi:10.48550/arXiv.2302.12173.

[147] Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles L Isbell, and Andrea L Thomaz. Policy shaping: Integrating human feedback with reinforcement learning. *Advances in neural information processing systems*, 26, 2013. URL: https://papers.nips.cc/paper_files/paper/2013/hash/e034fb6b66aacc1d48f445ddfb08da98-Abstract.html.

[148] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. BadNets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019. URL: https://ieeexplore.ieee.org/document/8685687, doi:10.1109/ACCESS.2019.2909068.

[149] Rachid Guerraoui, Arsany Guirguis, Jérémy Plassmann, Anton Ragot, and Sébastien Rouault. Garfield: System support for byzantine machine learning (regular paper). In *DSN*. IEEE, 2021. URL: https://arxiv.org/abs/2010.05888, doi:10.48550/arXiv.2010.05888.

[150] Chuan Guo, Alexandre Sablayrolles, Hervé Jégou, and Douwe Kiela. Gradient-based adversarial attacks against text transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5747–5757, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL: https://aclanthology.org/2021.emnlp-main.464, doi:10.18653/v1/2021.emnlp-main.464.

[151] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges, 2024. URL: https://arxiv.org/abs/2402.01680, arXiv:2402.01680, doi:10.48550/arXiv.2402.01680.

[152] Niv Haim, Gal Vardi, Gilad Yehudai, michal Irani, and Ohad Shamir. Reconstructing training data from trained neural networks. In Alice H. Oh, Alekh Agarwal, Danielle

美国，2021年。美国计算机协会。doi:10.1145/3490354.3494367.

[142] Shafi Goldwasser、Michael P.Kim、Vinod Vaikuntanathan 和 Or Zamir。在机器学习模型中植入不可检测的后门。https://arxiv.org/abs/2204.0 6974,2022。arXiv。doi:10.48550/ARXIV.2204.06974.

[143] Ian Goodfellow、Yoshua Bengio 和 Aaron Courville。深度学习。MIT Press,2016年。http://www.deeplearningbook.org.

[144] 伊恩·古德费洛、乔纳森·施伦斯和克里斯蒂安·塞格迪。解释与利用--对抗性样本。载于国际学习表征会议,2015年。网址：http://arxiv.org/abs/1412.6572, doi:10.48550/arXiv.1412.6572.

[145] Kai Greshake。提示注入防御应该更差一些。https://kai-greshake.de/posts/approaches-to-pi-defense/,2024年3月。访问日期：2024-08-22。

[146] Kai Greshake、Sahar Abdelnabi、Shailesh Mishra、Christoph Endres、Thorsten Holz 和 Mario Fritz。不是你注册时预期的情况：利用间接提示注入攻破现实世界中集成 LLM 的应用。arXiv 预印本 arXiv:2302.12173，2023 年。网址：https://arxiv.org/abs/2302.12173, doi:10.48550 /arXiv.2302.12173。

[147] Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles L Isbell，和 Andrea L Thomaz.策略塑造：将人类反馈与强化学习相结合。神经信息处理系统进展, 26, 2013. URL: https://papers.nips.cc/paper_files/paper/2013/hash/e034fb6b66aacc1d48f445ddfb08da98-Abs tract.html。

[148] Tianyu Gu、Kang Liu、Brendan Dolan-Gavitt 和 Siddharth Garg。BadNets：评估深度神经网络上的后门攻击。*IEEE Access*,7:47230-47244,2019。URL:https://ieeexplore.ieee.org/document/8685687, doi:10.1109/ACCESS.201 9.2909068。

[149] Rachid Guerraoui、Arsany Guirguis、Jérémy Plassmann、Anton Ragot 和 Sébastien Rouault。Garfield：拜占庭机器学习的系统支持(常规论文)。载于*DSN*。IEEE,2021。URL:https://arxiv.org/abs/2010.05888, doi:10.48550/arXiv.2010.05888。

[150] Chuan Guo、Alexandre Sablayrolles、Hervé Jégou 和 Douwe Kiela。针对文本 transformer 的基于梯度的对抗攻击。载于*2021* 年自然语言处理经验方法会议论文集，第 5747-5757 页，线上及多米尼加共和国蓬塔卡纳，2021 年 11 月。计算语言学协会。URL: https://aclanthology.org/2021.emnlp-main.464, doi:10.18653/v1/2021.emnlp-main.464.

[151] 郭泰成、陈秀英、王雅琪、常瑞迪、裴世超、Nitesh V.Chawla、Olaf Wiest 和张祥亮。基于大语言模型的多智能体：进展与挑战综述，2024。URL: https://arxiv.org/abs/2402.01680, arXiv:2402.01680, doi:10.48550/arXiv.2402.01680.

[152] Niv Haim、Gal Vardi、Gilad Yehudai、Michal Irani 和 Ohad Shamir。从训练好的神经网络中重建训练数据。载于 Alice H.Oh、Alekh Agarwal、Danielle

Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL: https://openreview.net/forum?id=Sxk8Bse3RKO, doi:10.48550/arXiv.2206.07758.

[153] Danny Halawi, Alexander Wei, Eric Wallace, Tony T. Wang, Nika Haghtalab, and Jacob Steinhardt. Covert malicious finetuning: Challenges in safeguarding llm adaptation, 2024. URL: https://arxiv.org/abs/2406.20053, arXiv:2406.20053, doi:10.48550/arXiv.2406.20053.

[154] Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. WildGuard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of LLMs, 2024. URL: https://arxiv.org/abs/2406.18495, arXiv:2406.18495, doi:10.48550/arXiv.2406.18495.

[155] Shanshan Han, Qifan Zhang, Yuhang Yao, Weizhao Jin, Zhaozhuo Xu, and Chaoyang He. LLM multi-agent systems: Challenges and open problems, 2024. URL: https://arxiv.org/abs/2402.03578, arXiv:2402.03578, doi:10.48550/arXiv.2402.03578.

[156] Drew Harwell. ID.me gathers lots of data besides face scans, including locations. Scammers still have found a way around it., retrieved December 2024. URL: https://www.washingtonpost.com/technology/2022/02/11/idme-facial-recognition-fraud-scams-irs/.

[157] Jonathan Hayase, Weihao Kong, Raghav Somani, and Sewoong Oh. SPECTRE: Defending against backdoor attacks using robust statistics. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4129–4139. PMLR, 18–24 Jul 2021. URL: https://proceedings.mlr.press/v139/hayase21a.html, doi:10.48550/arXiv.2104.11315.

[158] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ml safety, 2022. URL: https://arxiv.org/abs/2109.13916, arXiv:2109.13916, doi:10.48550/arXiv.2109.13916.

[159] Isaac Hepworth, Kara Olive, Kingshuk Dasgupta, Michael Le, Mark Lodato, Mihai Maruseac, Sarah Meiklejohn, Shamik Chaudhuri, and Tehila Minkus. Securing the ai software supply chain. Technical report, Google, 2024. URL: https://research.google/pubs/securing-the-ai-software-supply-chain/.

[160] Keegan Hines, Gary Lopez, Matthew Hall, Federico Zarfati, Yonatan Zunger, and Emre Kiciman. Defending against indirect prompt injection attacks with spotlighting, 2024. URL: https://arxiv.org/abs/2403.14720, arXiv:2403.14720, doi:10.48550/arXiv.2403.14720.

[161] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022. URL: https://arxiv.org/abs/2203.15556, arXiv:2203.15556.

Belgrave 和 Kyunghyun Cho 编辑，《神经信息处理系统进展》，2022年。网址：https://openreview.net/forum?id=Sxk8Bse3RKO, doi：10.48550/arXiv.2206.07758.

[153] Danny Halawi、Alexander Wei、Eric Wallace、Tony T.Wang、Nika Haghtalab 与 Jacob Steinhardt。隐蔽恶意微调：保障 LLM 适配中的挑战，2024。URL:https://arxiv.org/abs/2406.20053, arXiv:2406.20053,doi:10.485 50/arXiv.2406.20053。

[154] 韩承柱、卡维尔·拉奥、艾莉森·埃廷格、姜丽薇、林宇辰、内森·林bert、Yejin Choi 和 Nouha Dziri。WildGuard：面向 LLM 安全风险、越狱与拒答的开源一站式审核工具，2024。URL:https://arxiv.org/abs/2406.18495,arXiv:2406.18495,doi:10.48550/arXiv.2406.18495.

[155] Shanshan Han、Qifan Zhang、Yuhang Yao、Weizhao Jin、Zhaozhuo Xu 和 Chaoyang He。LLM 多智能体系统：挑战与开放问题，2024 年。URL:https://arxiv.org/abs/2402.03578,arXiv:2402.03578,doi:10.48550/arXiv.2402. 03578。

[156] @ @[156]@ @ Drew Harwell。ID.me 除面部扫描外还收集大量数据，包括位置信息。诈骗者仍找到了绕过它的方法。，检索于 2024 年 12 月。URL:https: //www.washingtonpost.com/technology/2022/02/11/idme-facial-recognition-fra ud-scams-irs/。

[157] 乔纳森·哈亚塞、孔伟豪、拉加夫·索马尼和吴承雄。SPECTRE：去-使用稳健统计防御后门攻击。在Marina Meila和Tong Zhang编辑，第38届国际机器学习会议论文集，机器学习研究论文集第139卷，第4129-4139页。PMLR，2021年7月18-24日。网址：https://proceedings.mlr.press/v139/hayase21a.html, doi:10.48550/arXiv.2104.11315.

[158] https://arxi Dan Hendrycks、Nicholas Carlini、John Schulman 和 Jacob Steinhardt。ML 安全中的未解决问题，2022 年。网址：⟨https://arxiv.org/abs/2109.13916,⟩ arXiv:2109.13916,doi:⟨10.48550/arXiv.2109.13916.⟩

[159] Isaac Hepworth、Kara Olive、Kingshuk Dasgupta、Michael Le、Mark Lodato、Mihai Maruseac、Sarah Meiklejohn、Shamik Chaudhuri 和 Tehila Minkus。保护 AI 软件供应链。技术报告，Google,2024 年。URL:https://research.googl e/pubs/securing-the-ai-software-supply-chain/。

[160] Keegan Hines、Gary Lopez、Matthew Hall、Federico Zarfati、Yonatan Zunger 和 Emre Kiciman。使用 spotlighting 防御间接提示注入攻击，2024。URL:https://arxiv.org/abs/2403.14720, arXiv:2403.14720,

doi:10.48550/arXiv.2403.14720。

[161] Jordan Hoffmann、Sebastian Borgeaud、Arthur Mensch、Elena Buchatskaya、Trevor Cai、Eliza Rutherford、Diego de Las Casas、Lisa Anne Hendricks、Johannes Welbl、Aidan Clark、Tom Hennigan、Eric Noland、Katie Millican、George van den Driessche、Bogdan Damoc、Aurelia Guy、Simon Osindero、Karen Simonyan、Erich Elsen、Jack W. Rae、Oriol Vinyals 和 Laurent Sifre。训练计算最优的大语言模型，2022年。网址：https://arxiv.org/abs/2203.15556, arXiv:2203.15556。

[162] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V Pearson, Dietrich A Stephan, Stanley F Nelson, and David W Craig. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS genetics*, 4(8):e1000167, 2008. URL: https://pubmed.ncbi.nlm.nih.gov/18769715/, doi:10.1371/journal.pgen.1000167.

[163] Xiaoling Hu, Xiao Lin, Michael Cogswell, Yi Yao, Susmit Jha, and Chao Chen. Trigger hunting with a topological prior for trojan detection. In *International Conference on Learning Representations*, 2022. URL: https://openreview.net/forum?id=TXsjU8BaibT, doi:10.48550/arXiv.2110.08335.

[164] Zhengmian Hu, Gang Wu, Saayan Mitra, Ruiyi Zhang, Tong Sun, Heng Huang, and Viswanathan Swaminathan. Token-level adversarial prompt detection based on perplexity measures and contextual information, 2024. URL: https://arxiv.org/abs/2311.11509, arXiv:2311.11509, doi:10.48550/arXiv.2311.11509.

[165] Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. Are large pre-trained language models leaking your personal information? *arXiv preprint arXiv:2205.12628*, 2022. URL: https://arxiv.org/abs/2205.12628, doi:10.48550/arXiv.2205.12628.

[166] W. Ronny Huang, Jonas Geiping, Liam Fowl, Gavin Taylor, and Tom Goldstein. Metapoison: Practical general-purpose clean-label data poisoning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12080–12091. Curran Associates, Inc., 2020. URL: https://proceedings.neurips.cc/paper/2020/file/8ce6fc704072e351679ac97d4a985574-Paper.pdf, doi:10.48550/arXiv.2004.00225.

[167] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents, 2022. URL: https://arxiv.org/abs/2201.07207, arXiv:2201.07207, doi:10.48550/arXiv.2201.07207.

[168] Xijie Huang, Moustafa Alzantot, and Mani Srivastava. NeuronInspect: Detecting backdoors in neural networks via output explanations, 2019. URL: https://arxiv.org/abs/1911.07399, doi:10.48550/ARXIV.1911.07399.

[169] Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Hanchi Sun, Zhengliang Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bertie Vidgen, Bhavya Kailkhura, Caiming Xiong, Chaowei Xiao, Chunyuan Li, Eric P. Xing, Furong Huang, Hao Liu, Heng Ji, Hongyi Wang, Huan Zhang, Huaxiu Yao, Manolis Kellis, Marinka Zitnik, Meng Jiang, Mohit Bansal, James Zou, Jian Pei, Jian Liu, Jianfeng Gao, Jiawei Han, Jieyu Zhao, Jiliang Tang, Jindong Wang, Joaquin Vanschoren, John Mitchell, Kai Shu, Kaidi Xu, Kai-Wei Chang, Lifang He, Lifu Huang, Michael Backes, Neil Zhenqiang Gong, Philip S. Yu, Pin-Yu Chen, Quanquan Gu, Ran Xu, Rex Ying, Shuiwang Ji, Suman Jana, Tianlong Chen, Tianming Liu, Tianyi Zhou, William Yang Wang, Xiang Li, Xiangliang Zhang, Xiao Wang, Xing Xie, Xun Chen, Xuyu Wang, Yan Liu, Yanfang Ye, Yinzhi Cao, Yong

[162] 尼尔斯·霍默、萨博尔奇·塞林格、玛戈特·雷德曼、大卫·杜根、瓦伊巴夫·坦贝、吉尔Muehling、John V Pearson、Dietrich A Stephan、Stanley F Nelson 和 David W Craig。使用高密度 SNP 基因分型微阵列解析向高度复杂混合物贡献微量 DNA 的个体。*PLoS genetics*,4(8):e1000167,2008. 网址：https://pubmed.ncbi.nlm.nih.gov/18769715/, doi:10.1371/journal.pgen.1000167。

[163] 胡晓玲、林晓、Michael Cogswell、姚毅、Susmit Jha 和 Chao Chen。基于拓扑先验的触发器搜寻用于木马检测。载于国际学习表征会议,2022年。网址：https://openreview.net/forum?id=TXsjU8BaibT，doi:10.48550/arXiv.2110.08335。

[164] 胡正棉、吴刚、Saayan Mitra、张瑞怡、孙彤、黄恒和Viswanathan Swaminathan。基于困惑度度量和上下文信息的Token级对抗性提示检测，2024年。URL: https://arxiv.org/abs/23 11.11509, arXiv:2311.11509, doi:10.48550/arXiv.2311.11509.

[165] Jie Huang、Hanyin Shao 和 Kevin Chen-Chuan Chang。大型预训练语言语言模型泄露你的个人信息?*arXiv*预印本*arXiv:2205.12628*,2022年。URL:https://arxiv.org/abs/2205.12628, doi:10.48550/arXiv.2205.1262

[166] W.Ronny Huang, Jonas Geiping, Liam Fowl, Gavin Taylor, and Tom Goldstein.Metapoison：实用的通用干净标签数据投毒。载于H.Larochelle, M.Ranzato, R.Hadsell, M.F.Balcan, and H.Lin编辑，*Advances in Neural Information Processing Systems*，第33卷，第12080-12091页。Curran Associates, Inc.,2020年。URL:https://proceedings.neurips.cc/paper/2020/file/8ce6fc704072e3516 79ac97d4a985574-Paper.pdf,doi:10.48550/arXiv.2004.00225.

[167] 黄文龙、彼得·阿比尔、迪帕克·帕塔克和伊戈尔·莫尔达奇。语言模型作为零样本规划器：为具身智能体提取可操作知识，2022年。URL: https://arxiv.org/abs/2201.07207, arXiv:2201.07207, doi: 10.48550/arXiv.2201.07207.

[168] https://arxi 黄希杰、Moustafa Alzantot 和 Mani Srivastava。NeuronInspect：通过输出解释检测神经网络中的后门，2019年。网址：《https://arxiv.or》g/abs/1911.07399,doi:《10.48550/ARXIV.1911.07399.》

[169] 黄悦、孙立超、王浩然、吴思远、张启辉、李源、高楚杰、黄奕欣、吕文瀚、张一璇、李欣然、孙涵驰、刘正亮、刘奕欣、王奕珏、张智坤、Bertie Vidgen、Bhavya Kailkhura、熊才明、肖朝伟、李春源、Eric P.Xing、黄芙蓉、刘浩、季恒、王弘毅、张焕、姚华旭、Manolis Kellis、Marinka Zitnik、姜猛、Mohit Bansal、James Zou、裴健、刘健、高剑锋、韩家炜、赵婕妤、唐继良、王劲东、Joaquin Vanschoren、John Mitchell、舒凯、徐凯迪、张凯威、何丽芳、黄立夫、Michael Backes、龚振强、Philip S.Yu、陈品谕、顾全全、徐然、Rex Ying、纪水旺、Suman Jana、陈天龙、刘天明、周天翼、William Yang Wang、李翔、张祥亮、王晓、谢幸、陈勋、王旭宇、刘燕、叶燕芳、曹寅之、Yong

Chen, and Yue Zhao. Position: TrustLLM: Trustworthiness in large language models. In *Forty-first International Conference on Machine Learning*, 2024. URL: https://openreview.net/forum?id=bWUU0LwwMp, doi:10.48550/arXiv.2401.05561.

[170] Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam Jermyn, Amanda Askell, Ansh Radhakrishnan, Cem Anil, David Duvenaud, Deep Ganguli, Fazl Barez, Jack Clark, Kamal Ndousse, Kshitij Sachan, Michael Sellitto, Mrinank Sharma, Nova DasSarma, Roger Grosse, Shauna Kravec, Yuntao Bai, Zachary Witten, Marina Favaro, Jan Brauner, Holden Karnofsky, Paul Christiano, Samuel R. Bowman, Logan Graham, Jared Kaplan, Sören Mindermann, Ryan Greenblatt, Buck Shlegeris, Nicholas Schiefer, and Ethan Perez. Sleeper agents: Training deceptive llms that persist through safety training, 2024. URL: https://arxiv.org/abs/2401.05566, arXiv:2401.05566, doi:10.48550/arXiv.2201.07207.

[171] W. Nicholson Price II. Risks and remedies for artificial intelligence in health care. https://www.brookings.edu/research/risks-and-remedies-for-artificial-intelligence-in-health-care/, 2019. Brookings Report.

[172] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2142–2151. PMLR, 2018. URL: http://proceedings.mlr.press/v80/ilyas18a.html, doi:10.48550/arXiv.1804.08598.

[173] Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Prior convictions: Black-box adversarial attacks with bandits and priors. In *International Conference on Learning Representations*, 2019. URL: https://openreview.net/forum?id=BkMiWhR5K7, doi:10.48550/arXiv.1807.07978.

[174] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/file/e2c420d928d4bf8ce0ff2ec19b371514-Paper.pdf, doi:10.48550/arXiv.1905.02175.

[175] Zachary Izzo, Mary Anne Smart, Kamalika Chaudhuri, and James Zou. Approximate data deletion from machine learning models. In Arindam Banerjee and Kenji Fukumizu, editors, *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pages 2008–2016. PMLR, 2021. URL: http://proceedings.mlr.press/v130/izzo21a.html, doi:10.48550/arXiv.2002.10077.

[176] Shahin Jabbari, Han-Ching Ou, Himabindu Lakkaraju, and Milind Tambe. An empirical study of the trade-offs between interpretability and fairness. In *ICML Workshop on Human Interpretability in Machine Learning, International Conference on Machine Learning (ICML)*, 2020. URL: https://teamcore.seas.harvard.edu/files/team

Chen, and Yue Zhao.Position: TrustLLM：大语言模型中的可信度。载于第四十一届国际机器学习会议,2024年。URL:https://op enreview.net/forum?id=bWUU0LwwMp,doi:10.48550/arXiv.2401.05561.

[170] Evan Hubinger、Carson Denison、Jesse Mu、Mike Lambert、Meg Tong、Monte MacDiarmid、Tamera Lanham、Daniel M. Ziegler、Tim Maxwell、Newton Cheng、Adam Jermyn、Amanda Askell、Ansh Radhakrishnan、Cem Anil、David Duvenaud、Deep Ganguli、Fazl Barez、Jack Clark、Kamal Ndousse、Kshitij Sachan、Michael Sellitto、Mrinank Sharma、Nova DasSarma、Roger Grosse、Shauna Kravec、Yuntao Bai、Zachary Witten、Marina Favaro、Jan Brauner、Holden Karnofsky、Paul Christiano、Samuel R. Bowman、Logan Graham、Jared Kaplan、Sören Mindermann、Ryan Greenblatt、Buck Shlegeris、Nicholas Schiefer 和 Ethan Perez。潜伏特工：训练能通过安全训练的欺骗性大型语言模型，2024年。网址：https://arxiv.org/abs/2401.05566, arXiv:2401.05566，doi:10.48550/arXiv.2201.07207.

[171] W.Nicholson Price II。人工智能在医疗保健中的风险与补救措施。电子医疗保健中的https://www.brookings.edu/research/risks-and-remedies-for-artificial-intelligenc,2019年。布鲁金斯报告。

[172] Andrew Ilyas、Logan Engstrom、Anish Athalye 和 Jessy Lin。黑盒对抗性攻有限查询和信息下的攻击。载于Jennifer G.Dy and Andreas Krause编辑，*Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*,*Proceedings of Machine Learning Research*第80卷，第2142-2151页。PMLR,2018年。URL:http://Proceedings.mlr.press/v80/ilyas18a.html,doi:10.48550/arXiv.1804.08598.

[173] Andrew Ilyas, Logan Engstrom, and Aleksander Madry.Prior convictions：基于bandits和先验的黑盒对抗攻击。载于*International Conference on Learning Representations*,2019年。URL:https://openreview.net/forum?id=BkMiWhR5K7, doi:10.48550/arXiv.1807.07978。

[174] Andrew Ilyas、Shibani Santurkar、Dimitris Tsipras、Logan Engstrom、Brandon Tran 和 Aleksander Madry。对抗样本不是缺陷，而是特征。载于 H. Wallach、H. Larochelle、A. Beygelzimer、F. d'Alché-Buc、E. Fox 和 R. Garnett 编，《神经信息处理系统进展》，第 32 卷。Curran Associates, Inc.，2019 年。URL：https://proceedings.neurips.cc/paper/2019/file/e2c420d 928d4bf8ce0ff2ec19b371514-Paper.pdf，doi：10.48550/arXiv.1905.02175.

[175] 扎卡里·伊佐、玛丽·安妮·斯马特、卡玛莉卡·乔杜里和詹姆斯·邹。近似-从机器学习模型中删除数据。载于Arindam Banerjee and Kenji Fukumizu编辑，*The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*,*Proceedings of Machine Learning Research*第130卷，第2008-2016页。PMLR,2021年。URL:http://Proceedings.mlr.press/v130/izzo21a.html,doi:10.48550/arXiv.2002.10077.

[176] Shahin Jabbari、Han-Ching Ou、Himabindu Lakkaraju 和 Milind Tambe。一项实证可解释性与公平性之间权衡的实证研究。载于*ICML Workshop on Human interpretability in Machine Learning, International Conference on Machine Learning (ICML)*,2020年。URL:https://teamcore.seas.harvard.edu/files/team

core/files/2020_jabbari_paper_32.pdf.

[177] Matthew Jagielski, Nicholas Carlini, David Berthelot, Alex Kurakin, and Nicolas Papernot. High accuracy and high fidelity extraction of neural networks. In *Proceedings of the 29th USENIX Conference on Security Symposium*, SEC'20, USA, 2020. USENIX Association. URL: https://arxiv.org/abs/1909.01838, doi:10.48550/arXiv.1909.01838.

[178] Matthew Jagielski, Michael Kearns, Jieming Mao, Alina Oprea, Aaron Roth, Saeed Sharifi Malvajerdi, and Jonathan Ullman. Differentially private fair learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 3000–3008. PMLR, 2019. URL: https://arxiv.org/abs/1812.02696, doi:10.48550/arXiv.1812.02696.

[179] Matthew Jagielski, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, and Bo Li. Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In *IEEE Symposium on Security and Privacy (S&P)*, pages 19–35, 2018. URL: https://arxiv.org/abs/1804.00308, doi:10.48550/arXiv.1804.00308.

[180] Matthew Jagielski, Giorgio Severi, Niklas Pousette Harger, and Alina Oprea. Subpopulation data poisoning attacks. In *Proceedings of the ACM Conference on Computer and Communications Security*, CCS, 2021. URL: https://arxiv.org/abs/2006.14026, doi:10.48550/arXiv.2006.14026.

[181] Matthew Jagielski, Jonathan Ullman, and Alina Oprea. Auditing differentially private machine learning: How private is private SGD? In *Advances in Neural Information Processing Systems*, volume 33, pages 22205–22216, 2020. URL: https://proceedings.neurips.cc/paper/2020/file/fc4ddc15f9f4b4b06ef7844d6bb53abf-Paper.pdf, doi:10.48550/arXiv.2006.07709.

[182] Neel Jain, Avi Schwarzschild, Yuxin Wen, Gowthami Somepalli, John Kirchenbauer, Ping yeh Chiang, Micah Goldblum, Aniruddha Saha, Jonas Geiping, and Tom Goldstein. Baseline defenses for adversarial attacks against aligned language models, 2023. URL: https://arxiv.org/abs/2309.00614, arXiv:2309.00614, doi:10.48550/arXiv.2309.00614.

[183] Bargav Jayaraman and David Evans. Evaluating differentially private machine learning in practice. In *Proceedings of the 28th USENIX Conference on Security Symposium*, SEC'19, page 1895–1912, USA, 2019. USENIX Association. URL: https://arxiv.org/abs/1902.08874, doi:10.48550/arXiv.1902.08874.

[184] Bargav Jayaraman and David Evans. Are attribute inference attacks just imputation? In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, CCS '22, page 1569–1582, New York, NY, USA, 2022. Association for Computing Machinery. doi:10.1145/3548606.3560663.

[185] Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark, September

core/files/2020_jabbari_paper_32.pdf。

[177] Matthew Jagielski、Nicholas Carlini、David Berthelot、Alex Kurakin 和 Nicolas Papernot。神经网络的高精度高保真提取。载于*Proceedings of the 29th USENIX Conference on Security Symposium*,SEC'20,USA,2020年。USENIX Association。URL:https://arxiv.org/abs/1909.01838, doi:10.48550/arXiv.190 9.01838。

[178] Matthew Jagielski、Michael Kearns、Jieming Mao、Alina Oprea、Aaron Roth、Saeed Sharifi Malvajerdi 和 Jonathan Ullman。差分隐私公平学习。载于 Kamalika Chaudhuri 和 Ruslan Salakhutdinov 编，《第 36 届国际机器学习大会论文集》，《机器学习研究论文集》，第 3000-3008 页。PMLR，2019 年。URL：https://arxiv.org/abs/1812.02696, doi:10.48550/arXiv.1812.02696。

[179] Matthew Jagielski, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, and Bo Li.操纵机器学习：回归学习的投毒攻击与对策。载于*IEEE Symposium on Security and Privacy (S&P)*，第19-35页，2018年。URL:https://arxiv.org/abs/1804.00308, doi:10.48550/arXiv.1804.00308.

[180] 马修·贾吉尔斯基、乔治奥·塞韦里、尼克拉斯·普塞特·哈格尔和艾琳娜·奥普雷亚。子群体-模拟数据投毒攻击。载于*Proceedings of the ACM Conference on Computer and Communications Security*,CCS,2021年。URL:https://arxiv.org/abs/2006.14026, doi:10.48550/arXiv.2006.14026。

[181] Matthew Jagielski, Jonathan Ullman, and Alina Oprea.审计差分隐私机器学习：私有SGD有多私密?载于*Advances in Neural Information Processing Systems*，第33卷，第22205-22216页，2020年。URL:https://proceedi ngs.neurips.cc/paper/2020/file/fc4ddc15f9f4b4b06ef7844d6bb53abf-Paper.pdf, doi:10.48550/arXiv.2006.07709。

[182] Neel Jain、Avi Schwarzschild、Yuxin Wen、Gowthami Somepalli、John Kirchenbauer、Ping yeh Chiang、Micah Goldblum、Aniruddha Saha、Jonas Geiping 和 Tom Goldstein。针对对齐语言模型对抗攻击的基线防御，2023。URL: https://arxiv.org/abs/2309.00614, arXiv:2309.00614, doi:10.48550/arXiv.2309.00614。

[183] Bargav Jayaraman 和 David Evans。评估差分隐私机器学习实践中的攻击。载于*Proceedings of the 28th USENIX Conference on Security Symposium*,SEC'19，第1895-1912页，USA,2019年。USENIX Association。URL:https://arxiv.org/abs/1902.08874,doi:10.48550/arXiv.1902.08874.

[184] Bargav Jayaraman and David Evans.属性推断攻击只是插补吗?载于*Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*,CCS '22，第1569-1582页，New York, NY, USA,2022年。Association for Computing Machinery。doi:10.1145/3548606.3560663.

[185] Robin Jia 和 Percy Liang。用于评估阅读理解能力的对抗样本系统。载于*Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*，第2021-2031页，Copenhagen, Denmark，九月

2017. Association for Computational Linguistics. URL: https://aclanthology.org/D17-1215, doi:10.18653/v1/D17-1215.

[186] Fengqing Jiang, Zhangchen Xu, Luyao Niu, Zhen Xiang, Bhaskar Ramasubramanian, Bo Li, and Radha Poovendran. Artprompt: Ascii art-based jailbreak attacks against aligned llms, 2024. URL: https://arxiv.org/abs/2402.11753, arXiv:2402.11753, doi:10.48550/arXiv.2402.11753.

[187] Pengfei Jing, Qiyi Tang, Yuefeng Du, Lei Xue, Xiapu Luo, Ting Wang, Sen Nie, and Shi Wu. Too good to be safe: Tricking lane detection in autonomous driving with crafted perturbations. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 3237–3254. USENIX Association, August 2021. URL: https://www.usenix.org/conference/usenixsecurity21/presentation/jing.

[188] Nikola Jovanovic, Robin Staab, and Martin Vechev. Watermark Stealing in Large Language Models. In *Proceedings of the 41-st International Conference on Machine Learning*, PMLR 235, June 2024. URL: https://files.sri.inf.ethz.ch/website/papers/jovanovic2024watermarkstealing.pdf, doi:10.48550/arXiv.2402.19361.

[189] Peter Kairouz, Brendan Mcmahan, Shuang Song, Om Thakkar, Abhradeep Thakurta, and Zheng Xu. Practical and private (deep) learning without sampling or shuffling. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5213–5225. PMLR, 18–24 Jul 2021. URL: https://proceedings.mlr.press/v139/kairouz21b.html.

[190] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning, 2019. URL: https://arxiv.org/abs/1912.04977, doi:10.48550/ARXIV.1912.04977.

[191] Guy Katz, Clark Barrett, David L. Dill, Kyle Julian, and Mykel J. Kochenderfer. Reluplex: An efficient SMT solver for verifying deep neural networks. In Rupak Majumdar and Viktor Kunčak, editors, *Computer Aided Verification*, pages 97–117, Cham, 2017. Springer International Publishing. URL: https://arxiv.org/abs/1702.01135, doi:10.48550/arXiv.1702.01135.

[192] Michael Kearns and Ming Li. Learning in the presence of malicious errors. In *Proceedings of the Twentieth Annual ACM Symposium on Theory of Computing*, STOC '88, page 267–280, New York, NY, USA, 1988. Association for Computing Machinery.

2017. 计算语言学协会。网址：https://aclanthology.org/D 17-1215,doi:10.18653/v1/D17-1215.

[186] Fengqing Jiang, Zhangchen Xu, Luyao Niu, Zhen Xiang, Bhaskar Ramasubramanian, Bo Li, and Radha Poovendran.Artprompt：基于ASCII艺术的针对对齐LLM的越狱攻击，2024年。URL: https://arxiv.org/abs/2402.11753, arXiv:2402.11753,

doi:10.48550/arXiv.2402.11753。

[187] Pengfei Jing, Qiyi Tang, Yuefeng Du, Lei Xue, Xiapu Luo, Ting Wang, Sen Nie, and Shi Wu.太好反而不安全：用精心构造的扰动欺骗自动驾驶中的车道检测。载于*30th USENIX Security Symposium (USENIX Security 21)*，第3237-3254页。USENIX Association,2021年8月。URL:https://www.usenix.org /conference/USENIX Security21/presentation/jing。

[188] Nikola Jovanovic, Robin Staab, and Martin Vechev.大语言模型中的水印窃取。载于*Proceedings of the 41-st International Conference on Machine Learning*,PMLR 235,2024年6月。URL:https://files.sri.inf.ethz.ch/website/papers/jovanovic2024Watermark Stealing.pdf,doi:10.48550/arXiv.2402.19361.

[189] Peter Kairouz, Brendan Mcmahan, Shuang Song, Om Thakkar, Abhradeep Thakurta, and Zheng Xu.无需采样或混洗的实用且私密的(深度)学习。载于Marina Meila and Tong Zhang编辑，*Proceedings of the 38th International Conference on Machine Learning*,*Proceedings of Machine Learning Research*第139卷，第5213-5225页。PMLR,2021年7月18-24日。URL:https://Proceedings.mlr.pr ess/v139/kairouz21b.html。

[190] 彼得·凯鲁兹，H. 布伦丹·麦克马汉，布伦丹·阿文特，奥雷利安·贝莱，迈赫迪·本尼斯，阿琼·尼廷·巴戈吉，卡利斯塔·博纳维茨，扎卡里·查尔斯，格雷厄姆·科莫德，雷切尔·卡明斯，拉斐尔·G。 L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. 斯蒂奇、齐腾·孙、阿南达·瑟塔·苏雷什、弗洛里安·特拉梅尔、普拉尼思·维帕科马、王建宇、李雄、徐政、杨强、费利克斯·X。 余、韩愈和森·赵。 联邦学习中的进展与开放问题，2019。 网址：https://arxiv.org/abs/1912.04977, doi:10.48550/ARXIV.1912.04977.

[191] Guy Katz, Clark Barrett, David L.Dill, Kyle Julian, and Mykel J.Kochenderfer.Reluplex：一种用于验证深度神经网络的高效SMT求解器。载于Rupak Majumdar and Viktor Kunčak编辑，*Computer Aided Verification*，第97-117页，Cham,2017年。Springer International Publishing。URL:https://arxiv.org/abs/1702.01135, doi:10.48550/arXiv.1702.01135.

[192] Michael Kearns 和 Ming Li。在存在恶意错误的情况下学习。载于*Pro-*
载于《第二十届 ACM 计算理论年会论文集》，STOC '88，第 267–280 页，美国纽约州纽约市，1988 年。美国计算机协会。

doi:10.1145/62212.62238.

[193] Alaa Khaddaj, Guillaume Leclerc, Aleksandar Makelov, Kristian Georgiev, Hadi Salman, Andrew Ilyas, and Aleksander Madry. Rethinking backdoor attacks. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 16216–16236. PMLR, 23–29 Jul 2023. URL: https://proceedings.mlr.press/v202/khaddaj23a.html, doi:10.48550/arXiv.2307.10163.

[194] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models, 2023. URL: https://arxiv.org/abs/2301.10226, arXiv:2301.10226, doi:10.48550/arXiv.2301.10226.

[195] Marius Kloft and Pavel Laskov. Security analysis of online centroid anomaly detection. *Journal of Machine Learning Research*, 13(118):3681–3724, 2012. URL: http://jmlr.org/papers/v13/kloft12b.html.

[196] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1885–1894. JMLR. org, 2017. URL: https://arxiv.org/abs/1703.04730, doi:10.48550/arXiv.1703.04730.

[197] Tomasz Korbak, Kejian Shi, Angelica Chen, Rasika Bhalerao, Christopher L. Buckley, Jason Phang, Samuel R. Bowman, and Ethan Perez. Pretraining language models with human preferences, 2023. URL: https://arxiv.org/abs/2302.08582, arXiv:2302.08582, doi:10.48550/arXiv.2302.08582.

[198] Moshe Kravchik, Battista Biggio, and Asaf Shabtai. Poisoning attacks on cyber attack detectors for industrial control systems. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, SAC '21, page 116–125, New York, NY, USA, 2021. Association for Computing Machinery. doi:10.1145/3412841.3441892.

[199] Ram Shankar Siva Kumar, Magnus Nyström, John Lambert, Andrew Marshall, Mario Goertzel, Andi Comissoneru, Matt Swann, and Sharon Xia. Adversarial machine learning – industry perspectives. https://arxiv.org/abs/2002.05646, 2020. doi:10.48550/ARXIV.2002.05646.

[200] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. https://arxiv.org/abs/1607.02533, 2016. doi:10.48550/ARXIV.1607.02533.

[201] Keita Kurita and Paul Michel amd Graham Neubig. Weight poisoning attacks on pretrained models, 2020. URL: https://arxiv.org/abs/2004.06660, arXiv:2004.06660, doi:10.48550/arXiv.2004.06660.

[202] E. La Malfa and M. Kwiatkowska. The king is naked: On the notion of robustness for natural language processing. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, volume 10, page 11047–57. Association for the Advancement of Artificial Intelligence, 2022. URL: https://arxiv.org/abs/2112.07605, doi:10.48550/arXiv.2112.07605.

[203] Ricky Laishram and Vir Virander Phoha. Curie: A method for protecting SVM classi-

doi:10.1145/62212.62238。

[193] Alaa Khaddaj、Guillaume Leclerc、Aleksandar Makelov、Kristian Georgiev、Hadi Salman、Andrew Ilyas 和 Aleksander Madry。重新思考后门攻击。载于 Andreas Krause、Emma Brunskill、Kyunghyun Cho、Barbara Engelhardt、Sivan Sabato 和 Jonathan Scarlett 编，《第40届国际机器学习会议论文集》，机器学习研究论文集第202卷，第16216-16236页。PMLR，2023年7月23-29日。网址：https://proceedings.mlr.press/v202/khaddaj23a.html，doi：10.48550/arXiv.2307.10163.

[194] John Kirchenbauer、Jonas Geiping、Yuxin Wen、Jonathan Katz、Ian Miers 和 Tom Goldstein。大型语言模型的水印，2023 年。网址：https://arxiv.org/abs/2301.10226,arXiv:2301.10226,doi:10.48550/arXiv.2301.10226.

[195] Marius Kloft 和 Pavel Laskov。在线质心异常检测的安全性分析检测。机器学习研究杂志， 13(118):3681-3724, 2012. URL: http://jmlr.org/papers/v13/kloft12b.html.

[196] Pang Wei Koh and Percy Liang.通过影响函数理解黑盒预测。载于*Proceedings of the 34th International Conference on Machine Learning-Volume 70*，第1885-1894页。JMLR.org,2017年。URL:https://arxiv.org/abs/1703.0 4730,doi:10.48550/arXiv.1703.04730.

[197] Tomasz Korbak、Kejian Shi、Angelica Chen、Rasika Bhalerao、Christopher L.Buckley、Jason Phang、Samuel R.Bowman 和 Ethan Perez。基于人类偏好的语言模型预训练，2023。URL:https://arxiv.org/abs/2302.08582, arXiv:2302.08582,doi:10.48550/arXiv.2302.08582.

[198] Moshe Kravchik, Battista Biggio, and Asaf Shabtai.针对工业控制系统网络攻击检测器的投毒攻击。载于*Proceedings of the 36th Annual ACM Symposium on Applied Computing*,SAC '21，第116-125页，New York, NY, USA,2021年。Association for Computing Machinery。doi:10.1145/3412841.3441892.

[199] Ram Shankar Siva Kumar、Magnus Nyström、John Lambert、Andrew Marshall、Mario Goertzel、Andi Comissoneru、Matt Swann 和 Sharon Xia。对抗性机器学习--行业视角。https://arxiv.org/abs/2002.05646, 2020。doi:10.48550/ARXIV.2002.05646。

[200] Alexey Kurakin、Ian Goodfellow 和 Samy Bengio。物理世界中的对抗样本--现实世界。https://arxiv.org/abs/1607.02533, 2016 年。doi：10.48550/ARXIV.1607.02533。

[201] Keita Kurita 和 Paul Michel 以及 Graham Neubig。针对预训练模型的权重投毒攻击--经过训练的模型，2020年。网址：https://arxiv.org/abs/2004.06660, arXiv:2004.06660,doi:10.48550/arXiv.2004.06660.

[202] E.La Malfa and M.Kwiatkowska.国王没穿衣服：论自然语言处理的鲁棒性概念。载于*Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*，第10卷，第11047-57页。Association for the Advancement of Artificial Intelligence,2022年。URL:https://arxiv.org/abs/2112.07605, doi:10.48550/arXiv.2112.07605.

[203] Ricky Laishram 和 Vir Virander Phoha。Curie：一种保护SVM分类器

fier from poisoning attack. *CoRR*, abs/1606.01584, 2016. URL: http://arxiv.org/abs/1606.01584, arXiv:1606.01584, doi:10.48550/arXiv.1606.01584.

[204] Lakera. Guard, 2023. URL: https://www.lakera.ai/.

[205] Harry Langford, Ilia Shumailov, Yiren Zhao, Robert D. Mullins, and Nicolas Papernot. Architectural neural backdoors from first principles. *CoRR*, abs/2402.06957, 2024. URL: https://doi.org/10.48550/arXiv.2402.06957, arXiv:2402.06957, doi:10.48550/ARXIV.2402.06957.

[206] Learn Prompting. Defensive measures, 2023. URL: https://learnprompting.org/docs/category/-defensive-measures.

[207] Mathias Lécuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19-23, 2019*, pages 656–672. IEEE, 2019. doi:10.1109/SP.2019.00044.

[208] Klas Leino and Matt Fredrikson. Stolen memories: Leveraging model memorization for calibrated white-box membership inference. In *Proceedings of the 29th USENIX Conference on Security Symposium*, SEC'20, USA, 2020. USENIX Association. URL: https://arxiv.org/abs/1906.11798, doi:10.48550/arXiv.1906.11798.

[209] Alexander Levine and Soheil Feizi. Deep partition aggregation: Provable defenses against general poisoning attacks. In *International Conference on Learning Representations*, 2021. URL: https://openreview.net/forum?id=YUGG2tFuPM, doi:10.48550/arXiv.2006.14768.

[210] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. arXiv:2005.11401.

[211] Linyi Li, Tao Xie, and Bo Li. Sok: Certified robustness for deep neural networks. In *44th IEEE Symposium on Security and Privacy, SP 2023, San Francisco, CA, USA, 22-26 May 2023*. IEEE, 2023. URL: https://arxiv.org/abs/2009.04131, doi:10.48550/arXiv.2009.04131.

[212] Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Adam Khoja, Zhenqi Zhao, Ariel Herbert-Voss, Cort B. Breuer, Samuel Marks, Oam Patel, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Lin, Adam A. Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Russell Kaplan, Ian Steneker, David Campbell, Brad Jokubaitis, Alex Levinson, Jean Wang, William Qian, Kallol Krishna Karmakar, Steven Basart, Stephen Fitz, Mindy Levine, Ponnurangam Kumaraguru, Uday Tupakula, Vijay Varadharajan, Ruoyu Wang, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. The wmdp benchmark: Measuring and reducing malicious use with unlearning, 2024. URL: https://arxiv.org/abs/2403.03218, arXiv:2403.03218, doi:

免受投毒攻击的方法。*CoRR*,abs/1606.01584,2016。URL:http://arxiv.org/abs/1606.01584,arXiv:1606.01584,doi:10.48550/arXiv.1606.01584.

[204] Lakera。Guard，2023 年。URL：https://www.lakera.ai/.

[205] Harry Langford、Ilia Shumailov、Yiren Zhao、Robert D.Mullins 和 Nicolas Papernot。基于第一性原理的架构神经后门。*CoRR*,abs/2402.06957,2024。URL:https://doi.org/10.48550/arXiv.2402.06957, arXiv:2402.06957,doi:10.48550/ARXIV.2402.06957.

[206] Learn Prompting。防御措施，2023。网址：https://Learn Prompting.org/docs/category/-防御措施。

[207] Mathias Lécuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana.具有差分隐私的对抗样本认证鲁棒性。载于*2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19-23, 2019* ，第656-672页。IEEE,2019年。doi:10.1109/SP.2019.00044.

[208] Klas Leino and Matt Fredrikson.被窃取的记忆：利用模型记忆进行校准的白盒成员推断。载于*Proceedings of the 29th USENIX Conference on Security Symposium*,SEC'20,USA,2020年。USENIX Association。URL:https://arxiv.org/abs/1906.11798, doi:10.48550/arXiv.1906.11798.

[209] Alexander Levine and Soheil Feizi.深度划分聚合：针对一般投毒攻击的可证明防御。载于*International Conference on Learning Representations*,2021年。URL:https://openreview.net/forum?id=YUGG2tFuPM, doi:10.48550/arXiv.2006.14768。

[210] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela.面向知识密集型nlp任务的检索增强生成， 2021. arXiv:2005.11401.

[211] Linyi Li, Tao Xie, and Bo Li.Sok：深度神经网络的认证鲁棒性。载于*44th IEEE Symposium on Security and Privacy, SP 2023, San Francisco, CA, USA, 22-26 May 2023*年5月26日.IEEE,2023。网址：https://arxiv.org/abs/2009.04131, doi:10.48550 / arXiv.2009.04131。

[212] Nathaniel Li、Alexander Pan、Anjali Gopal、Summer Yue、Daniel Berrios、Alice Gatti、Justin D. Li、Ann-Kathrin Dombrowski、Shashwat Goel、Long Phan、Gabriel Mukobi、Nathan Helm-Burger、Rassin Lababidi、Lennart Justen、Andrew B. Liu、Michael Chen、Isabelle Barrass、Oliver Zhang、Xiaoyuan Zhu、Rishub Tamirisa、Bhrugu Bharathi、Adam Khoja、Zhenqi Zhao、Ariel Herbert-Voss、Cort B. Breuer、Samuel Marks、Oam Patel、Andy Zou、Mantas Mazeika、Zifan Wang、Palash Oswal、Weiran Lin、Adam A. Hunt、Justin Tienken-Harder、Kevin Y. Shih、Kemper Talley、John Guan、Russell Kaplan、Ian Steneker、David Campbell、Brad Jokubaitis、Alex Levinson、Jean Wang、William Qian、Kallol Krishna Karmakar、Steven Basart、Stephen Fitz、Mindy Levine、Ponnurangam Kumaraguru、Uday Tupakula、Vijay Varadharajan、Ruoyu Wang、Yan Shoshitaishvili、Jimmy Ba、Kevin M. Esvelt、Alexandr Wang 和 Dan Hendrycks。WMDP 基准：通过遗忘来测量和减少恶意使用，2024年。网址：https://arxiv.org/abs/2403.03218, arXiv:2403.03218，doi:

10.48550/arXiv.2403.03218.

[213] Shaofeng Li, Hui Liu, Tian Dong, Benjamin Zi Hao Zhao, Minhui Xue, Haojin Zhu, and Jialiang Lu. Hidden backdoors in human-centric language models. In Yongdae Kim, Jong Kim, Giovanni Vigna, and Elaine Shi, editors, *CCS '21: 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, Republic of Korea, November 15 - 19, 2021*, pages 3123–3140. ACM, 2021. doi:10.1145/3460120.3484576.

[214] Shaofeng Li, Minhui Xue, Benjamin Zi Hao Zhao, Haojin Zhu, and Xinpeng Zhang. Invisible backdoor attacks on deep neural networks via steganography and regularization. *IEEE Transactions on Dependable and Secure Computing*, 18:2088–2105, 2021. URL: https://arxiv.org/abs/1909.02742, doi:10.48550/arXiv.1909.02742.

[215] Shasha Li, Ajaya Neupane, Sujoy Paul, Chengyu Song, Srikanth V. Krishnamurthy, Amit K. Roy-Chowdhury, and Ananthram Swami. Adversarial perturbations against real-time video classification systems. *CoRR*, abs/1807.00458, 2018. URL: http://arxiv.org/abs/1807.00458, arXiv:1807.00458, doi:10.14722/ndss.2019.23202.

[216] Ji Lin, Chuang Gan, and Song Han. Defensive quantization: When efficiency meets robustness. *ArXiv*, abs/1904.08444, 2019. URL: https://api.semanticscholar.org/CorpusID:53502621, doi:10.48550/arXiv.1904.08444.

[217] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In Michael Bailey, Sotiris Ioannidis, Manolis Stamatogiannakis, and Thorsten Holz, editors, *Research in Attacks, Intrusions, and Defenses - 21st International Symposium, RAID 2018, Proceedings*, Lecture Notes in Computer Science, pages 273–294. Springer Verlag, 2018. URL: https://link.springer.com/chapter/10.1007/978-3-030-00470-5_13, doi:10.1007/978-3-030-00470-5_13.

[218] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *International Conference on Learning Representations*, 2017. URL: https://openreview.net/forum?id=Sys6GJqxl, doi:10.48550/arXiv.1611.02770.

[219] Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, and Yang Liu. Prompt injection attack against LLM-integrated applications. *arXiv preprint arXiv:2306.05499*, 2023. URL: https://arxiv.org/abs/2306.05499, doi:10.48550/arXiv.2306.05499.

[220] Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, and Yang Liu. Jailbreaking ChatGPT via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860*, 2023. URL: https://arxiv.org/abs/2305.13860, doi:10.48550/arXiv.2305.13860.

[221] Yingqi Liu, Wen-Chuan Lee, Guanhong Tao, Shiqing Ma, Yousra Aafer, and Xiangyu Zhang. ABS: Scanning neural networks for back-doors by artificial brain stimulation. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, CCS '19, page 1265–1282, New York, NY, USA, 2019. Association for Computing Machinery. doi:10.1145/3319535.3363216.

10.48550/arXiv.2403.03218.

[213] 李少峰、刘辉、董天、Benjamin Zi Hao Zhao、薛敏慧、朱浩瑾和陆佳亮。以人为中心的语言模型中的隐藏后门。载于 Yongdae Kim、Jong Kim、Giovanni Vigna 和 Elaine Shi 编，*CCS '21:2021 ACM SIGSAC* 计算机与通信安全会议，虚拟活动，韩国，*2021*年*11*月*15*日至*19*日，第3123-3140页。ACM,2021年。doi:10.1145/34 60120.3484576。

[214] 李少峰、薛敏慧、赵秉浩、朱浩金、张新鹏。在-通过隐写术和正则化对深度神经网络的可见后门攻击。*IEEE Transactions on Dependable and Secure Computing*, 18:2088-2105, 2021. URL: https://arxiv.org/abs/1909.02742, doi:10.48550/arXiv.1909.02742.

[215] Shasha Li、Ajaya Neupane、Sujoy Paul、Chengyu Song、Srikanth V.Krishnamurthy、Amit K.Roy-Chowdhury 和 Ananthram Swami。针对实时视频分类系统的对抗性扰动。*CoRR*,abs/1807.00458,2018。URL:http://ar xiv.org/abs/1807.00458,arXiv:1807.00458,doi:10.14722/ndss.2019.23202.

[216] Ji Lin, Chuang Gan, and Song Han.防御性量化：当效率遇上鲁棒性。*ArXiv*, abs/1904.08444, 2019. URL: https://api.semanticscholar.org/ CorpusID:53502621, doi:10.48550/arXiv.1904.08444.

[217] Kang Liu、Brendan Dolan-Gavitt 和 Siddharth Garg。精细剪枝：防御深度神经网络后门攻击。载于 Michael Bailey、Sotiris Ioannidis、Manolis Stamatogiannakis 和 Thorsten Holz 编，攻击、入侵与防御研究--第*21*届国际研讨会 *RAID 2018* 论文集，计算机科学讲义，第273-294页。Springer Verlag,2018年。URL: https://link.springer.com/chapter/10.1007/978-3-030-00470-5_13,

doi:10.1007/978-3-030-00470-5_13。

[218] 刘延培、陈欣云、刘畅和宋晓东。深入探究可迁移的对抗对抗样本与黑盒攻击。载于国际学习表征会议,2017年。网址：https://openreview.net/forum?id=Sys6GJqxl,

[219] Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, and Yang Liu.针对 LLM 集成应用的提示注入攻击。*arXiv preprint arXiv:2306.05499*, 2023. URL: https://arxiv.org/abs/23 06.05499, doi:10.48550/arXiv.2306.05499.

[220] @ @[220]@ @ Yi Liu、Gelei Deng、Zhengzi Xu、Yuekang Li、Yaowen Zheng、Ying Zhang、Lida Zhao、Tianwei Zhang 和 Yang Liu。通过提示工程越狱 ChatGPT：一项实证研究。*arXiv* 预印本 *arXiv:2305.13860*,2023 年。网址：@ @https://arxiv.org/abs/23@ @ 05.13860,doi:@ @10.48550/arXiv.2305.13860.@ @

[221] Yingqi Liu、Wen-Chuan Lee、Guanhong Tao、Shiqing Ma、Yousra Aafer 和 Xiangyu Zhang。ABS：通过人工大脑刺激扫描神经网络中的后门。载于*2019* 年 *ACM SIGSAC* 计算机与通信安全会议论文集,CCS '19，第 1265-1282 页，美国纽约州纽约市，2019 年。美国计算机协会。doi:10.1145/3319535.3363216.

[222] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *NDSS*. The Internet Society, 2018. URL: http://dblp.uni-trier.de/db/conf/ndss/ndss2018.html#LiuMALZW018.

[223] Yunfei Liu, Xingjun Ma, James Bailey, and Feng Lu. Reflection backdoor: A natural backdoor attack on deep neural networks. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 182–199, Cham, 2020. Springer International Publishing. URL: https://arxiv.org/abs/2007.02343, doi:10.48550/arXiv.2007.02343.

[224] Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, and toxicity, 2023. URL: https://arxiv.org/abs/2305.13169, arXiv:2305.13169, doi:10.48550/arXiv.2305.13169.

[225] Martin Bertran Lopez, Shuai Tang, Michael Kearns, Jamie Morgenstern, Aaron Roth, and Zhiwei Steven Wu. Scalable membership inference attacks via quantile regression. In *NeurIPS 2023*, 2023. URL: https://www.amazon.science/publications/scalable-membership-inference-attacks-via-quantile-regression.

[226] Daniel Lowd and Christopher Meek. Adversarial learning. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05, page 641–647, New York, NY, USA, 2005. Association for Computing Machinery. doi:10.1145/1081870.1081950.

[227] Jiajun Lu, Hussein Sibai, Evan Fabry, and David Forsyth. No need to worry about adversarial examples in object detection in autonomous vehicles, 2017. URL: https://arxiv.org/abs/1707.03501, arXiv:1707.03501, doi:10.48550/arXiv.1707.03501.

[228] Yiwei Lu, Gautam Kamath, and Yaoliang Yu. Indiscriminate data poisoning attacks on neural networks. https://arxiv.org/abs/2204.09092, 2022. doi:10.48550/ARXIV.2204.09092.

[229] Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. Analyzing leakage of personally identifiable information in language models. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 346–363. IEEE Computer Society, 2023. URL: https://arxiv.org/abs/2302.00539, doi:10.48550/arXiv.2302.00539.

[230] Yuzhe Ma, Xiaojin Zhu, and Justin Hsu. Data poisoning against differentially-private learners: Attacks and defenses. In *In Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, 2019.

[231] Pooria Madani and Natalija Vlajic. Robustness of deep autoencoder in intrusion detection under adversarial contamination. In *HoTSoS '18: Proceedings of the 5th Annual Symposium and Bootcamp on Hot Topics in the Science of Security*, pages 1–8, 04 2018. doi:10.1145/3190619.3190637.

[232] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In

[222] 刘英奇、马世清、Yousra Aafer、李文川、翟娟、王伟航，以及张翔宇。对神经网络的木马攻击。载于*NDSS*。互联网协会，2018. 网址：http://dblp.uni-trier.de/db/conf/ndss/ndss2018.html#LiuMALZW018.

[223] 刘云飞、马兴军、James Bailey 和卢峰。反射后门：一种自然深度神经网络中的后门攻击。载于 Andrea Vedaldi、Horst Bischof、Thomas Brox 和 Jan-Michael Frahm 编，*Computer Vision - ECCV 2020*，第 182-199 页，Cham,2020。Springer International Publishing。URL:https://arxiv.org/ab

[224] Shayne Longpre、Gregory Yauney、Emily Reif、Katherine Lee、Adam Roberts、Barret Zoph、Denny Zhou、Jason Wei、Kevin Robinson、David Mimno 和 Daphne Ippolito。预训练者的训练数据指南：衡量数据年龄、领域覆盖、质量和毒性的影响，2023。URL: https://arxiv.org/abs/2305.13169, arXiv:2305.13169, doi:10.48550/arXiv.2305.13169.

[225] Martin Bertran Lopez、Shuai Tang、Michael Kearns、Jamie Morgenstern、Aaron Roth 和 Zhiwei Steven Wu。通过分位数回归的可扩展成员推断攻击。载于*NeurIPS 2023*,2023。URL: https://www.amazon.science/publications/scal able-membership-inference-attacks-via-quantile-regression。

[226] Daniel Lowd and Christopher Meek.对抗学习.In 第十一届 *ACM SIGKDD* 知识发现与数据挖掘国际会议论文集, KDD '05， 第 641-647 页， 美国纽约州纽约市， 2005. 美国计算机协会.doi:10.1145/1081870.1081950.

[227] Jiajun Lu、Hussein Sibai、Evan Fabry 和 David Forsyth。无需担心自动驾驶车辆中目标检测的对抗样本问题，2017。URL:https://arxiv.org/abs/1707.03501,arXiv:1707.03501,doi:10.48550/arXiv.1707. 03501。

[228] Yiwei Lu、Gautam Kamath 和 Yaoliang Yu。针对神经网络的不可区分数据投毒(poisoning)攻击。https://arxiv.org/abs/2204.09092, 2022。doi:10.48550/ARX IV.2204.09092。

[229] Nils Lukas、Ahmed Salem、Robert Sim、Shruti Tople、Lukas Wutschitz 和 Santiago Zanella-Béguelin。分析语言模型中的个人可识别信息泄露。载于*2023 IEEE Symposium on Security and Privacy (SP)*，第 346-363 页。IEEE Computer Society,2023。URL:https://arxiv.org/abs/2302.00539,

[230] Yuzhe Ma、Xiaojin Zhu 和 Justin Hsu。针对差分隐私学习者的数据投毒：攻击与防御。载于*In Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*,2019。

[231] Pooria Madani 和 Natalija Vlajic。对抗污染下深度自编码器在入侵检测中的鲁棒性。载于*HoTSoS '18: Proceedings of the 5th Annual Symposium and Bootcamp on Hot Topics in the Science of Security*，第 1-8 页，04 2018。doi:10.1145/3190619.3190637.

[232] Aleksander Madry、Aleksandar Makelov、Ludwig Schmidt、Dimitris Tsipras 和 Adrian Vladu。迈向抵御对抗攻击的深度学习模型。载于

6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net, 2018. URL: https://openreview.net/forum?id=rJzIBfZAb, doi:10.48550/arXiv.1706.06083.

[233] Saeed Mahloujifar, Esha Ghosh, and Melissa Chase. Property inference from poisoning. In *2022 IEEE Symposium on Security and Privacy (S&P)*, pages 1120–1137, 2022. URL: https://ieeexplore.ieee.org/document/9833623, doi:10.1109/SP46214.2022.9833623.

[234] Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. TOFU: A task of fictitious unlearning for LLMs, 2024. URL: https://arxiv.org/abs/2401.06121, arXiv:2401.06121, doi:10.48550/arXiv.2401.06121.

[235] Neal Mangaokar, Ashish Hooda, Jihye Choi, Shreyas Chandrashekaran, Kassem Fawaz, Somesh Jha, and Atul Prakash. Prp: Propagating universal perturbations to attack large language model guard-rails, 2024. URL: https://arxiv.org/abs/2402.15911, arXiv:2402.15911, doi:10.48550/arXiv.2402.15911.

[236] James Manyika and Sissie Hsiao. An overview of Bard: an early experiment with generative AI. https://ai.google/static/documents/google-about-bard.pdf, February 2023. Google.

[237] Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal, 2024. URL: https://arxiv.org/abs/2402.04249, arXiv:2402.04249, doi:10.48550/arXiv.2402.15911.

[238] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *IEEE Symposium on Foundations of Computer Science*, FOCS '07, pages 94–103, Las Vegas, NV, USA, 2007. URL: https://ieeexplore.ieee.org/document/4389483, doi:10.1109/FOCS.2007.66.

[239] Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. Tree of attacks: Jailbreaking black-box LLMs automatically. *arXiv preprint arXiv:2312.02119*, 2023. URL: https://arxiv.org/abs/2312.02119, doi:10.48550/arXiv.2312.02119.

[240] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19-23, 2019*, pages 691–706. IEEE, 2019. doi:10.1109/SP.2019.00029.

[241] Melissa Heikkilä. This new data poisoning tool lets artists fight back against generative AI. https://www.technologyreview.com/2023/10/23/1082189/data-poisoning-artists-fight-generative-ai/, October 2023. MIT Technology Review.

[242] El Mahdi El Mhamdi, Sadegh Farhadkhani, Rachid Guerraoui, Arsany Guirguis, Lê-Nguyên Hoang, and Sébastien Rouault. Collaborative learning in the jungle (decentralized, byzantine, heterogeneous, asynchronous and nonconvex learning). In *NeurIPS*, 2021. URL: https://arxiv.org/abs/2008.00742, doi:10.48550/arXiv.2

第六届国际学习表征会议，*ICLR 2018*，加拿大不列颠哥伦比亚省温哥华，*2018*年*4*月*30*日至*5*月*3*日，会议论文集。OpenReview.net,
2018. 网址：https://openreview.net/forum?id=rJzIBfZAb, doi：10.48550/arXiv.1706.06083。

[233] [222] Saeed Mahloujifar、Esha Ghosh 和 Melissa Chase。来自 poi- 的属性推断推理。在2022年IEEE安全与隐私研讨会（S&P）上，第1120-1137页，2022年。网址：https://ieeexplore.ieee.org/document/9833623, doi：10.1109/SP46 214.2022.9833623。

[234] Pratyush Maini、Zhili Feng、Avi Schwarzschild、Zachary C.Lipton 和 J.Zico Kolter。TOFU：面向 LLM 的虚构遗忘任务，2024。URL:https://arxiv.org/abs/2401.06121,arXiv:2401.06121,doi:10.48550/arXiv.2401.06121.

[235] Neal Mangaokar、Ashish Hooda、Jihye Choi、Shreyas Chandrashekaran、Kassem Fawaz、Somesh Jha 和 Atul Prakash。Prp：传播通用扰动以攻击大型语言模型护栏，2024。URL:https://arxiv.org/abs/2402. 15911,arXiv:2402.15911,doi:10.48550/arXiv.2402.15911.

[236] James Manyika 和 Sissie Hsiao。Bard 概述：生成式 AI 的早期实验。https://ai.google/static/documents/google-about-bard.pdf, 2023 年 2 月。Google。

[237] Mantas Mazeika、Long Phan、Xuwang Yin、Andy Zou、Zifan Wang、Norman Mu、Elham Sakhaee、Nathaniel Li、Steven Basart、Bo Li、David Forsyth 和 Dan Hendrycks。Harmbench：用于自动化红队测试与稳健拒绝的标准化评估框架，2024 年。URL: https://arxiv.org/abs/2402.04249, arXiv:2402.04249,
doi:10.48550/arXiv.2402.15911。

[238] Frank McSherry 和 Kunal Talwar。通过差分隐私的机制设计。载于*IEEE Symposium on Foundations of Computer Science*,FOCS '07，第 94-103 页，Las Vegas, NV, USA,2007。URL:https://ieeexplore.ieee.org/document/4389483,
doi:10.1109/FOCS.2007.66。

[239] 阿奈·梅赫罗特拉、马诺利斯·赞佩塔基斯、保罗·卡西亚尼克、布莱恩·纳尔逊、海勒姆·安德-son、Yaron Singer 和 Amin Karbasi。攻击树：自动越狱黑盒 LLM。*arXiv preprint arXiv:2312.02119*,2023。URL:https://arxiv.org/abs/ 2312.02119,doi:10.48550/arXiv.2312.02119.

[240] Luca Melis、Congzheng Song、Emiliano De Cristofaro和Vitaly Shmatikov。利用协作学习中的意外特征泄露。2019年IEEE安全与隐私研讨会，SP 2019，美国加利福尼亚州旧金山，2019年5月19-23日，第691-706页。IEEE，2019。doi:10.1109/SP.2019.00029.

[241] 梅丽莎·海基拉。这款新的数据投毒工具让艺术家们得以反击生成-生成式人工智能。https://www.Technology Review.com/2023/10/23/1082189/data-poisoni ng-artists-fight-generative-ai/,2023年10月。MIT Technology Review。

[242] El Mahdi El Mhamdi、Sadegh Farhadkhani、Rachid Guerraoui、Arsany Guirguis、Lê-Nguyên Hoang 和 Sébastien Rouault。丛林中的协作学习(去中心化、拜占庭、异构、异步和非凸学习)。载于*NeurIPS*,2021。URL:https://arxiv.org/abs/2008.00742, doi:10.48550/arXiv.2

008.00742.

[243] El Mahdi El Mhamdi, Rachid Guerraoui, and Sébastien Rouault. The Hidden Vulnerability of Distributed Learning in Byzantium. In *ICML*, 2018. URL: https://arxiv.org/abs/1802.07927, doi:10.48550/arXiv.1802.07927.

[244] El Mahdi El Mhamdi, Rachid Guerraoui, and Sébastien Rouault. Distributed momentum for byzantine-resilient stochastic gradient descent. In *ICLR*, 2021. URL: https://arxiv.org/abs/2003.00010, doi:10.48550/arXiv.2003.00010.

[245] Dang Minh, H. Xiang Wang, Y. Fen Li, and Tan N. Nguyen. You can't have AI both ways: Balancing health data privacy and access fairly. *Artificial Intelligence Review volume*, 55:3503–3568, 2022. https://doi.org/10.1007/s10462-021-10088-y. doi:10.3389/fgene.2022.929453.

[246] Ilya Mironov, Kunal Talwar, and Li Zhang. R\'enyi differential privacy of the sampled gaussian mechanism. *arXiv preprint arXiv:1908.10530*, 2019. URL: https://arxiv.org/abs/1908.10530, doi:10.48550/arXiv.1908.10530.

[247] Margaret Mitchell, Giada Pistilli, Yacine Jernite, Ezinwanne Ozoani, Marissa Gerchick, Nazneen Rajani, Sasha Luccioni, Irene Solaiman, Maraim Masoud, Somaieh Nikpoor, Carlos Muñoz Ferrandis, Stas Bekman, Christopher Akiki, Danish Contractor, David Lansky, Angelina McMillan-Major, Tristan Thrush, Suzana Ilić, Gérard Dupont, Shayne Longpre, Manan Dey, Stella Biderman, Douwe Kiela, Emi Baylor, Teven Le Scao, Aaron Gokaslan, Julien Launay, and Niklas Muennighoff. BigScience Large Open-science Open-access Multilingual Language Model. https://huggingface.co/bigscience/bloom, 2022. Hugging Face.

[248] MITRE. ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems, retrieved December 2024. URL: https://atlas.mitre.org/.

[249] MITRE ATLAS. AML.M0001: Limit Model Artifact Release. https://atlas.mitre.org/mitigations/AML.M0001, 2023. Last Modified: 12 October 2023.

[250] MITRE ATLAS. AML.M0002: Passive ML Output Obfuscation. https://atlas.mitre.org/mitigations/AML.M0002, 2023. Last Modified: 12 October 2023.

[251] MITRE ATLAS. AML.M0004: Restrict Number of ML Model Queries. https://atlas.mitre.org/mitigations/AML.M0004, 2023. Last Modified: 12 October 2023.

[252] MITRE ATLAS. AML.M0000: Limit Release of Public Information. https://atlas.mitre.org/mitigations/AML.M0000, 2024. Last Modified: 12 January 2024.

[253] Payman Mohassel and Yupeng Zhang. SecureML: A system for scalable privacy-preserving machine learning. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 19–38, 2017. doi:10.1109/SP.2017.12.

[254] Seungyong Moon, Gaon An, and Hyun Oh Song. Parsimonious black-box adversarial attacks via efficient combinatorial optimization. In *International Conference on Machine Learning (ICML)*, 2019. URL: https://arxiv.org/abs/1905.06635, doi:10.48550/arXiv.1905.06635.

[255] Olivia Moore. How Are Consumers Using Generative AI? *Andreessen Horowitz (a16z)*, 2023. URL: https://a16z.com/how-are-consumers-using-generative-ai/.

[256] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard.

008.00742.

[243] El Mahdi El Mhamdi、Rachid Guerraoui 和 Sébastien Rouault。隐藏的脆弱性--拜占庭环境下分布式学习的能力.In ICML, 2018. URL: https://arxiv.org/ abs/1802.07927, doi:10.48550/arXiv.1802.07927.

[244] El Mahdi El Mhamdi、Rachid Guerraoui 和 Sébastien Rouault。分布式动量法用于拜占庭鲁棒随机梯度下降。载于*ICLR*,2021年。URL:https://arxiv.org/abs/2003.00010, doi:10.48550/arXiv.2003.00010.

[245] Dang Minh、H. Xiang Wang、Y. Fen Li和Tan N. Nguyen。你不能两全其美地使用AI：公平地平衡健康数据隐私与访问。人工智能评论卷，55：3503-3568，2022年。https://doi.org/10.1007/s10462-021-10088-y.

doi:10.3389/fgene.2022.929453。

[246] Ilya Mironov, Kunal Talwar, and Li Zhang.采样高斯机制的 R \ 'enyi 差分隐私.*arXiv* 预印本 *arXiv:1908.10530*, 2019. URL: https://arxiv.or g/abs/1908.10530, doi:10.48550/arXiv.1908.10530.

[247] Margaret Mitchell, Giada Pistilli, Yacine Jernite, Ezinwanne Ozoani, Marissa Gerchick, Nazneen Rajani, Sasha Luccioni, Irene Solaiman, Maraim Masoud, Somaieh Nikpoor, Carlos Muñoz Ferrandis, Stas Bekman, Christopher Akiki, Danish Contractor, David Lansky, Angelina McMillan-Major, Tristan Thrush, Suzana Ilić, Gérard Dupont, Shayne Longpre, Manan Dey, Stella Biderman, Douwe Kiela, Emi Baylor, Teven Le Scao, Aaron Gokaslan, Julien Launay， 和 Niklas Muennighoff。BigScience 大型开放科学开放获取多语言语言模型。https://huggingfac e.co/bigscience/bloom, 2022。Hugging Face。

[248] MITRE。ATLAS：人工智能系统的对抗性威胁态势，re-检索于2024年12月。网址：https://atlas.mitre.org/.

[249] [222] MITRE ATLAS。AML.M0001：限制模型产物的发布。(https://atlas.mitre.org/) 缓解措施/AML.M0001,2023年。最后修改日期：2023年10月12日。

[250] [222] MITRE ATLAS。AML.M0002：被动式机器学习输出混淆。(https://atlas.mitre.or) g/mitigations/AML.M0002,2023年。最后修改日期：2023年10月12日。

[251] MITRE ATLAS。AML.M0004：限制 ML 模型查询次数。https://atlas.mitre.org/mitigations/AML.M0004,2023年。最后修改日期：2023年10月12日。

[252] MITRE ATLAS。AML.M0000：限制公开信息的发布。https://atlas.mitre.org/mitigations/AML.M0000,2024年。最后修改日期：2024年1月12日。

[253] Payman Mohassel and Yupeng Zhang.SecureML：一个可扩展的隐私保护机器学习系统.In *2017 IEEE* 安全与隐私研讨会*(SP)*，第 19-38 页，2017. doi:10.1109/SP.2017.12.

[254] 文承勇、安佳温、宋贤五。简约黑盒对抗性通过高效组合优化实现的对抗攻击。载于国际机器学习会议*(ICML)*,2019年。网址：https://arxiv.org/abs/1905.06635,

[255] Olivia Moore。消费者如何使用生成式人工智能?*Andreessen Horowitz (a16z)*,2023。URL:https://a16z.com/how-are-consumers-using-Generative-ai/.

[256] Seyed-Mohsen Moosavi-Dezfooli、Alhussein Fawzi、Omar Fawzi和Pascal Frossard。

Universal adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. URL: https://arxiv.org/abs/1610.08401, doi:10.48550/arXiv.1610.08401.

[257] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. DeepFool: A simple and accurate method to fool deep neural networks. https://arxiv.org/abs/1511.04599, 2015. doi:10.48550/ARXIV.1511.04599.

[258] Ghulam Muhammad, Fatima Alshehri, Fakhri Karray, Abdulmotaleb El Saddik, Mansour Alsulaiman, and Tiago H. Falk. A comprehensive survey on multimodal medical signals fusion for smart healthcare systems. *Information Fusion*, 76:355–375, 2021. URL: https://www.sciencedirect.com/science/article/pii/S1566253521001330, doi:10.1016/j.inffus.2021.06.007.

[259] Sasi Kumar Murakonda and Reza Shokri. ML Privacy Meter: Aiding regulatory compliance by quantifying the privacy risks of machine learning, 2020. URL: https://arxiv.org/abs/2007.09339, doi:10.48550/ARXIV.2007.09339.

[260] Luis Muñoz-González, Battista Biggio, Ambra Demontis, Andrea Paudice, Vasin Wongrassamee, Emil C. Lupu, and Fabio Roli. Towards poisoning of deep learning algorithms with back-gradient optimization. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, AISec '17, 2017. URL: https://arxiv.org/abs/1708.08689, doi:10.48550/arXiv.1708.08689.

[261] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. Webgpt: Browser-assisted question-answering with human feedback, 2022. URL: https://arxiv.org/abs/2112.09332, arXiv:2112.09332, doi:10.48550/arXiv.2112.09332.

[262] Nina Narodytska and Shiva Kasiviswanathan. Simple black-box adversarial attacks on deep neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1310–1318, 2017. URL: https://ieeexplore.ieee.org/document/8014906, doi:10.1109/CVPRW.2017.172.

[263] Milad Nasr, Jamie Hayes, Thomas Steinke, Borja Balle, Florian Tramèr, Matthew Jagielski, Nicholas Carlini, and Andreas Terzis. Tight auditing of differentially private machine learning. In Joseph A. Calandrino and Carmela Troncoso, editors, *32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023*, pages 1631–1648. USENIX Association, 2023. URL: https://www.usenix.org/conference/usenixsecurity23/presentation/nasr, doi:10.48550/arXiv.2302.07956.

[264] Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *IEEE Symposium on Security and Privacy*, pages 739–753. IEEE, 2019. URL: https://arxiv.org/abs/1812.00910, doi:10.1109/SP.2019.00065.

[265] Milad Nasr, Shuang Song, Abhradeep Thakurta, Nicolas Papernot, and Nicholas

通用对抗扰动。载于《IEEE计算机视觉与模式识别会议（CVPR）论文集》，2017年7月。网址：https://arxiv.org/abs/ 1610.08401，doi：10.48550/arXiv.1610.08401.

[257] Seyed-Mohsen Moosavi-Dezfooli、Alhussein Fawzi 和 Pascal Frossard。DeepFool：一种简单而准确的欺骗深度神经网络的方法。https://arxiv.org/ab s/1511.04599,2015年。doi:10.48550/ARXIV.1511.04599.

[258] Ghulam Muhammad、Fatima Alshehri、Fakhri Karray、Abdulmotaleb El Saddik、Man- sour Alsulaiman 和 Tiago H.Falk。面向智能医疗系统的多模态医学信号融合综合综述。*Information Fusion*,76:355-375,2021年。URL:https://www.sciencedirect.com/science/article/pii/S1566253521001330,

doi:10.1016/j.inffus.2021.06.007。

[259] Sasi Kumar Murakonda 和 Reza Shokri。ML Privacy Meter：通过量化机器学习的隐私风险来辅助监管合

通过量化机器学习的隐私风险来实现合规，2020年。网址：https: //arxiv.org/abs/2007.09339，doi：10.48550/ARXIV.2007.09339.

[260] 路易斯·穆尼奥斯-冈萨雷斯，巴蒂斯塔·比吉奥，安布拉·德蒙蒂斯，安德烈亚·保迪切，瓦辛·旺-

grassamee、Emil C.Lupu 和 Fabio Roli。利用反向梯度优化对深度学习算法进行投毒。载于*Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*,AISec '17,2017。URL:https://arxiv.org/abs/ 1708.08689,doi:10.48550/arXiv.1708.08689.

[261] Reiichiro Nakano、Jacob Hilton、Suchir Balaji、Jeff Wu、Long Ouyang、Christina Kim、Christopher Hesse、Shantanu Jain、Vineet Kosaraju、William Saunders、Xu Jiang、Karl Cobbe、Tyna Eloundou、Gretchen Krueger、Kevin Button、Matthew Knight、Benjamin Chess和John Schulman。Webgpt：带人类反馈的浏览器辅助问答，2022年。网址：https://arxiv.org/abs/2112.09332, arXiv：2112.09332，

doi:10.48550/arXiv.2112.09332。

[262] Nina Narodytska 和 Shiva Kasiviswanathan。针对深度神经网络的简单黑盒对抗攻击。载于*2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*，第 1310-1318 页，2017。URL:https://ieeexplore.ieee.org/document/8014906,doi:10.1109/CVPRW.2017.172.

[263] Milad Nasr、Jamie Hayes、Thomas Steinke、Borja Balle、Florian Tramèr、Matthew Jagielski、Nicholas Carlini 和 Andreas Terzis。差分隐私机器学习的严格审计。载于 Joseph A. Calandrino 和 Carmela Troncoso 编，第32届 USENIX 安全研讨会，USENIX Security 2023，美国加利福尼亚州阿纳海姆，2023年8月9-11日，第1631-1648页。USENIX 协会，2023年。URL：https://www.usenix.org / conference/usenixsecurity23/presentation/nasr，doi：10.48550/arXiv.2302. 07956。

[264] Milad Nasr、Reza Shokri 和 Amir Houmansadr。深度学习的全面隐私分析：针对集中式和联邦学习的被动与主动白盒推断攻击。载于*IEEE Symposium on Security and Privacy*，第 739-753 页。IEEE,2019。URL:https://arxiv.org/abs/1812.00910, doi:10.1109/SP.2019.0006

5.

[265] Milad Nasr、Shuang Song、Abhradeep Thakurta、Nicolas Papernot 和 Nicholas

Carlini. Adversary instantiation: Lower bounds for differentially private machine learning. In *IEEE Symposium on Security & Privacy*, IEEE S&P '21, 2021. https://arxiv.org/abs/2101.04535. doi:10.48550/arXiv.2101.04535.

[266] National Cyber Security Centre. Machine learning principles. Technical report, National Cyber Security Centre, United Kingdom, 2024. Accessed: July 18, 2024. URL: https://www.ncsc.gov.uk/collection/machine-learning-principles.

[267] National Security Commission on Artificial Intelligence. Final report. https://www.nscai.gov/2021-final-report/, 2021. doi:10.48550/ARXIV.2006.03463.

[268] Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-delete: Gradient-based methods for machine unlearning. In Vitaly Feldman, Katrina Ligett, and Sivan Sabato, editors, *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pages 931–962. PMLR, 16–19 Mar 2021. URL: https://proceedings.mlr.press/v132/neel21a.html, doi:10.48550/arXiv.2007.02923.

[269] Blaine Nelson, Marco Barreno, Fuching Jack Chi, Anthony D. Joseph, Benjamin I.P. Rubinstein, Udam Saini, Charles Sutton, and Kai Xia. Exploiting machine learning to subvert your spam filter. In *First USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET 08)*, San Francisco, CA, April 2008. USENIX Association. URL: https://www.usenix.org/conference/leet-08/exploiting-machine-learning-subvert-your-spam-filter.

[270] J. Newsome, B. Karp, and D. Song. Polygraph: Automatically generating signatures for polymorphic worms. In *2005 IEEE Symposium on Security and Privacy (S&P)*, pages 226–241, 2005. URL: https://ieeexplore.ieee.org/document/1425070, doi:10.1109/SP.2005.15.

[271] Thien Duc Nguyen, Phillip Rieger, Huili Chen, Hossein Yalame, Helen Möllering, Hossein Fereidooni, Samuel Marchal, Markus Miettinen, Azalia Mirhoseini, Shaza Zeitouni, Farinaz Koushanfar, Ahmad-Reza Sadeghi, and Thomas Schneider. FLAME: Taming backdoors in federated learning. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 1415–1432, Boston, MA, August 2022. USENIX Association. URL: https://www.usenix.org/conference/usenixsecurity22/presentation/nguyen.

[272] Nisos. Building Trustworthy AI: Contending with Data Poisoning, retrieved December 2024. URL: https://www.nisos.com/research/building-trustworthy-ai/.

[273] NIST. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. https://airc.nist.gov/docs/NIST.AI.600-1.GenAI-Profile.ipd.pdf, 2024. NIST AI Publication (NIST AI) NIST AI 600-1 Initial Public Draft, National Institute of Standards and Technology, Gaitherburg, MD. doi:10.6028/NIST.AI.600-1.

[274] National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). https://doi.org/10.6028/NIST.AI.100-1, 2023. Online.

[275] National Institute of Standards and Technology. Managing misuse risk for dual-use

Carlini。对手实例化：差分隐私机器学习的下界。载于*IEEE Symposium on Security & Privacy*,IEEE S&P '21,2021。https: //arxiv.org/abs/2101.04535。doi:10.48550/arXiv.2101.04535.

[266] 英国国家网络安全中心。机器学习原则。技术报告，National Cyber Security Centre，英国，2024年。访问日期：2024年7月18日。URL:https://www.ncsc.gov.uk/collection/machine-learning-principles.

[267] 美国人工智能国家安全委员会。最终报告。https://www.nscai.gov/2021-final-report/，2021年。doi：10.48550/ARXIV.2006.03463.

[268] Seth Neel、Aaron Roth 和 Saeed Sharifi-Malvajerdi。Descent-to-delete：基于梯度的机器遗忘方法。载于 Vitaly Feldman、Katrina Ligett 和 Sivan Sabato 编，*Proceedings of the 32nd International Conference on Algorithmic Learning Theory*，第 132 卷，*Proceedings of Machine Learning Research*，第 931-962 页。PMLR,16-19 Mar 2021。URL:https://Proceedings.mlr.press/v132/neel21a.html，doi：10.48550/arXiv.2007.02923。

[269] Blaine Nelson、Marco Barreno、Fuching Jack Chi、Anthony D.Joseph、Benjamin I.P.Rubinstein、Udam Saini、Charles Sutton 和 Kai Xia。利用机器学习颠覆你的垃圾邮件过滤器。载于*First USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET 08)*,San Francisco, CA,April 2008。USENIX Association。URL:https://www.usenix.org/conference/leet-08/Exploiting-machine-learning-subvert - your-spam-filter。

[270] J.Newsome、B.Karp 和 D.Song。Polygraph：自动为多态蠕虫生成签名。载于*2005 IEEE Symposium on Security and Privacy (S&P)*，第 226-241 页，2005。URL:https://ieeexplore.ieee.org/document/1425070,

doi:10.1109/SP.2005.15。

[271] Thien Duc Nguyen, Phillip Rieger, Huili Chen, Hossein Yalame, Helen Möllering, Hossein Fereidooni, Samuel Marchal, Markus Miettinen, Azalia Mirhoseini, Shaza Zeitouni, Farinaz Koushanfar, Ahmad-Reza Sadeghi， 和 Thomas Schneider. FLAME：驯服联邦学习中的后门。载于第31届USENIX安全研讨会（USENIX Security 22），第1415-1432页，马萨诸塞州波士顿，2022年8月。USENIX协会。网址：https://www.usenix.org/conference/usenixsecurity22/presentation/ nguyen。

[272] Nisos。构建可信人工智能：应对数据投毒，检索于12月
2024年。URL: https://www.nisos.com/research/building-trustworthy-ai/.

[273] NIST.人工智能风险管理框架：生成式人工智能
概况。https://airc.nist.gov/docs/NIST.AI.600-1.GenAI-Profile.ipd.pdf, 2024. NIST AI 出版物(NIST AI)NIST AI 600-1 初始公开草案，美国国家标准与技术研究院，马里兰州盖瑟斯堡。doi:10.6028/NIST.AI.60 0-1。

[274] 美国国家标准与技术研究院。人工智能风险管理
管理框架（AI RMF 1.0）。https://doi.org/10.6028/NIST.AI.100-1, 2023年。在线。

[275] 美国国家标准与技术研究院。管理两用

foundation models. https://doi.org/10.6028/NIST.AI.800-1.ipd, 2024. Online.

[276] Parmy Olson. Faces are the next target for fraudsters, retrieved December 2024. URL: https://www.wsj.com/articles/faces-are-the-next-target-for-fraudsters-11625662828.

[277] OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[278] OpenAI. Assistants api overview, 2024. Accessed: 2024-08-18. URL: https://platform.openai.com/docs/assistants/overview.

[279] OpenAI. GPT-4o System Card. Online: https://cdn.openai.com/gpt-4o-system-card.pdf, August 2024.

[280] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Knockoff nets: Stealing functionality of black-box models. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4949–4958, 2019. URL: https://ieeexplore.ieee.org/document/8953839, doi:10.1109/CVPR.2019.00509.

[281] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL: https://openreview.net/forum?id=TG8KACxEON, doi:10.48550/arXiv.2203.02155.

[282] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. https://arxiv.org/abs/1605.07277, 2016. doi:10.48550/ARXIV.1605.07277.

[283] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, ASIA CCS '17, page 506–519, New York, NY, USA, 2017. Association for Computing Machinery. doi:10.1145/3052973.3053009.

[284] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy (S&P)*, pages 582–597, 2016. URL: https://ieeexplore.ieee.org/document/7546524, doi:10.1109/SP.2016.41.

[285] V. Pareto. *Manuale di Economia Politica*. Società Editrice Libraria, Milan, 1906.

[286] V. Pareto. *Manual of Political Economy*. Augustus M. Kelley Publishers, New York, 1971. URL: https://www.loc.gov/item/05022672/.

[287] Dario Pasquini, Martin Strohmeier, and Carmela Troncoso. Neural Exec: Learning (and learning from) execution triggers for prompt injection attacks, 2024. URL: https://arxiv.org/abs/2403.03792, arXiv:2403.03792, doi:10.48550/arXiv.2403.03792.

[288] Arpita Patra, Thomas Schneider, Ajith Suresh, and Hossein Yalame. ABY2.0: Improved Mixed-Protocol secure Two-Party computation. In *30th USENIX Security*

基础模型。https://doi.org/10.6028/NIST.AI.800-1.ipd, 2024。在线。

[276] Parmy Olson。人脸是欺诈者的下一个目标，检索于 2024 年 12 月。网址：https://www.wsj.com/articles/faces-are-the-next-target-for-fraudsters-11625662828。

[277] OpenAI。GPT-4 技术报告。*arXiv preprint arXiv:2303.08774*,2023。

[278] OpenAI。Assistants API 概述，2024。访问日期：2024-08-18。网址：https://platform.openai.com/docs/assistants/overview。

[279] OpenAI。GPT-4o 系统卡。在线：https://cdn.openai.com/gpt-4o-system-card.pdf,2024年8月。

[280] Tribhuvanesh Orekondy、Bernt Schiele 和 Mario Fritz。Knockoff 网络：窃取黑盒模型的功能性。在2019年IEEE/CVF计算机视觉与模式识别会议（CVPR），第4949-4958页，2019年。网址：https://ieeexplore.ieee. org/document/8953839，doi：10.1109/CVPR.2019.00509.

[281] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike， 和 Ryan Lowe。通过人类反馈训练语言模型遵循指令。载于 Alice H. Oh, Alekh Agarwal, Danielle Belgrave， 和 Kyunghyun Cho 编，Advances in Neural Information Processing Systems, 2022。URL: https://openreview.net/forum?id=TG8KACxEON,

doi:10.48550/arXiv.2203.02155。

[282] Nicolas Papernot、Patrick McDaniel 和 Ian Goodfellow。机器学习中的可迁移性：从现象到使用对抗样本的黑盒攻击。https://arxiv.org/abs/1605.07277,2016。doi:10.48550/ARXIV.1605.07277.

[283] Nicolas Papernot、Patrick McDaniel、Ian Goodfellow、Somesh Jha、Z.Berkay Celik 和 Ananthram Swami。针对机器学习的实用黑盒攻击。载于*Proceedings of the 2017 ACM on Asia Conference on Computer and Communica- tions Security*,ASIA CCS '17，第 506-519 页，New York, NY, USA,2017。Association for Computing Machinery。doi:10.1145/3052973.3053009.

[284] Nicolas Papernot、Patrick McDaniel、Xi Wu、Somesh Jha和Ananthram Swami。蒸馏作为防御深度神经网络对抗扰动的防御。在2016年IEEE安全与隐私研讨会（S&P），第582-597页，2016年。网址：https://ieeexplore.ieee.org/document/7546524, doi：10.1109/SP.2016.41.

[285] V.帕累托。政治经济学手册。Società Editrice Libraria，米兰，1906年。

[286] V.帕累托。政治经济学手册。Augustus M.Kelley 出版社，纽约，1971年。URL: https://www.loc.gov/item/05022672/.

[287] Dario Pasquini、Martin Strohmeier和Carmela Troncoso。Neural Exec：学习(并学习自)用于提示注入攻击的执行触发器，2024。URL:https://arxiv.org/abs/2403.03792,arXiv:2403.03792,doi:10.48550/arXiv.2403. 03792。

[288] Arpita Patra, Thomas Schneider, Ajith Suresh， 和 Hossein Yalame。ABY2.0： 实证明了混合协议安全的两方计算。载于第*30*届 *USENIX* 安全研讨会

Symposium (USENIX Security 21), pages 2165–2182. USENIX Association, August 2021. URL: https://www.usenix.org/conference/usenixsecurity21/presentat ion/patra.

[289] Andrea Paudice, Luis Muñoz-González, and Emil C. Lupu. Label sanitization against label flipping poisoning attacks. In Carlos Alzate, Anna Monreale, Haytham Assem, Albert Bifet, Teodora Sandra Buda, Bora Caglayan, Brett Drury, Eva García-Martín, Ricard Gavaldà, Stefan Kramer, Niklas Lavesson, Michael Madden, Ian Molloy, Maria-Irina Nicolae, and Mathieu Sinn, editors, *Nemesis/UrbReas/SoGood/IWAISe/GDM@PKDD/ECML*, volume 11329 of *Lecture Notes in Computer Science*, pages 5–15. Springer, 2018. URL: http://dblp.uni-trier.de/db/conf/pkdd/nem esis2018.html#PaudiceML18.

[290] Hammond Pearce, Baleegh Ahmad, Benjamin Tan, Brendan Dolan-Gavitt, and Ramesh Karri. Asleep at the keyboard? assessing the security of github copilot's code contributions, 2021. URL: https://arxiv.org/abs/2108.09293, arXiv: 2108.09293, doi:10.48550/arXiv.2108.09293.

[291] R. Perdisci, D. Dagon, Wenke Lee, P. Fogla, and M. Sharif. Misleading worm signature generators using deliberate noise injection. In *2006 IEEE Symposium on Security and Privacy (S&P'06)*, Berkeley/Oakland, CA, 2006. IEEE. URL: http://ieeexplore.ieee.or g/document/1623998/, doi:10.1109/SP.2006.26.

[292] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*, 2022. URL: https://arxiv. org/abs/2202.03286, doi:10.48550/arXiv.2202.03286.

[293] Neehar Peri, Neal Gupta, W. Ronny Huang, Liam Fowl, Chen Zhu, Soheil Feizi, Tom Goldstein, and John P. Dickerson. Deep k-nn defense against clean-label data poisoning attacks. In Adrien Bartoli and Andrea Fusiello, editors, *Computer Vision – ECCV 2020 Workshops*, pages 55–70, Cham, 2020. Springer International Publishing. URL: https://arxiv.org/abs/1909.13374, doi:10.48550/arXiv.1909.13374.

[294] Neil Perry, Megha Srivastava, Deepak Kumar, and Dan Boneh. Do users write more insecure code with ai assistants? In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, CCS '23. ACM, November 2023. URL: http://dx.doi.org/10.1145/3576915.3623157, doi:10.1145/3576915.3623157.

[295] Fabio Pierazzi, Feargus Pendlebury, Jacopo Cortellazzi, and Lorenzo Cavallaro. Intriguing properties of adversarial ML attacks in the problem space. In *2020 IEEE Symposium on Security and Privacy (S&P)*, pages 1308–1325. IEEE Computer Society, 2020. URL: https://doi.ieeecomputersociety.org/10.1109/SP40000.2020.00073, doi:10.1109/SP40000.2020.00073.

[296] Julien Piet, Maha Alrashed, Chawin Sitawarin, Sizhe Chen, Zeming Wei, Elizabeth Sun, Basel Alomair, and David Wagner. Jatmo: Prompt injection defense by task-specific finetuning, 2024. URL: https://arxiv.org/abs/2312.17673, arXiv:2312.1 7673, doi:10.48550/arXiv.2312.17673.

[297] Krishna Pillutla, Galen Andrew, Peter Kairouz, H. Brendan McMahan, Alina Oprea,

研讨会*(USENIX Security 21)*，第2165-2182页。USENIX 协会，2021年8月。URL: https://www.usenix.org/conference/USENIX Security21/presentat ion/patra。

[289] Andrea Paudice, Luis Muñoz-González， 和 Emil C. Lupu。针对标签翻转投毒攻击的标签净化。载于 Carlos Alzate, Anna Monreale, Haytham Assem, Albert Bifet, Teodora Sandra Buda, Bora Caglayan, Brett Drury, Eva GarcíaMartín, Ricard Gavaldà, Stefan Kramer, Niklas Lavesson, Michael Madden, Ian Molloy, Maria-Irina Nicolae， 和 Mathieu Sinn 编，Nemesis/UrbReas/SoGood/IWAISe/GDM@PKDD/ECML，Lecture Notes in Computer Science 第 11329 卷，第 5-15 页。Springer, 2018。URL: http://dblp.uni-trier.de/db/conf/pkdd/nem esis2018.html#PaudiceML18。

[290] Hammond Pearce、Baleegh Ahmad、Benjamin Tan、Brendan Dolan-Gavitt 和 Ramesh Karri。键盘前睡着了吗?评估 GitHub Copilot 代码贡献的安全性，2021年。网址：https://arxiv.org/abs/2108.09293, arXiv: 2108.09293,doi:10.48550/arXiv.2108.09293.

[291] R.Perdisci、D.Dagon、Wenke Lee、P.Fogla 和 M.Sharif。利用故意噪声注入误导蠕虫特征生成器。载于*2006*年 *IEEE* 安全与隐私研讨会*(S&P'06)*，伯克利/奥克兰，加利福尼亚州，2006年。IEEE。URL: http://ieeexplore.ieee.or g/document/1623998/,doi:10.1109/SP.2006.26.

[292] Ethan Perez、Saffron Huang、Francis Song、Trevor Cai、Roman Ring、John Aslanides、Amelia Glaese、Nat McAleese 和 Geoffrey Irving。用语言模型对语言模型进行红队测试。*arXiv* 预印本 *arXiv:2202.03286*,2022年。URL: https://arxiv.org/abs/2202.03286,doi:10.48550/arXiv.2202.03286.

[293] Neehar Peri、Neal Gupta、W.Ronny Huang、Liam Fowl、Chen Zhu、Soheil Feizi、Tom Goldstein 和 John P.Dickerson。针对干净标签数据投毒攻击的深度 k-nn 防御。载于 Adrien Bartoli 和 Andrea Fusiello 编，计算机视觉 - *ECCV 2020* 研讨会，第55-70页，Cham,2020年。Springer International Publishing。URL: https://arxiv.org/abs/1909.13374, doi:10.48550/arXiv.1909.13374.

[294] Neil Perry、Megha Srivastava、Deepak Kumar 和 Dan Boneh。用户使用 ai 助手会编写更多不安全的代码吗?载于*2023*年 *ACM SIGSAC* 计算机与通信安全会议论文集,CCS '23。ACM,2023年11月。URL: http://dx.doi.org/10.1145/3576915.3623157, doi:10.1145/3576915.3623157.

[295] Fabio Pierazzi、Feargus Pendlebury、Jacopo Cortellazzi和Lorenzo Cavallaro。In-问题空间中对抗性机器学习攻击的有趣特性。载于*2020*年 *IEEE* 安全与隐私研讨会*(S&P)*，第1308-1325页。IEEE 计算机学会，2020年。URL: https://doi.IEEE Computer Society.org/10.1109/SP40000.2020.00073,

[296] Julien Piet、Maha Alrashed、Chawin Sitawarin、Sizhe Chen、Zeming Wei、Elizabeth Sun、Basel Alomair 和 David Wagner。Jatmo：通过任务特定微调实现提示注入防御，2024。URL:https://arxiv.org/abs/2312.17673, arXiv:2312.1 7673,doi:10.48550/arXiv.2312.17673.

[297] 克里希纳·皮卢特拉，盖伦·安德鲁，彼得·凯鲁兹，H。布伦丹·麦克马汉，阿丽娜·奥普雷亚，

and Sewoong Oh. Unleashing the power of randomization in auditing differentially private ml. In *Advances in Neural Information Processing Systems*, 2023. URL: https://arxiv.org/abs/2305.18447, doi:10.48550/arXiv.2305.18447.

[298] PromptArmor and Kai Greshake. Data exfiltration from writer.com with indirect prompt injection, 2023. URL: https://promptarmor.substack.com/p/data-exfiltration-from-writercom.

[299] Jonathan Protzenko, Bryan Parno, Aymeric Fromherz, Chris Hawblitzel, Marina Polubelova, Karthikeyan Bhargavan, Benjamin Beurdouche, Joonwon Choi, Antoine Delignat-Lavaud, Cédric Fournet, Natalia Kulatova, Tahina Ramananandro, Aseem Rastogi, Nikhil Swamy, Christoph Wintersteiger, and Santiago Zanella-Beguelin. EverCrypt: A fast, verified, cross-platform cryptographic provider. In *Proceedings of the IEEE Symposium on Security and Privacy (Oakland)*, May 2020. URL: https://eprint.iacr.org/2019/757, doi:10.1109/SP40000.2020.00114.

[300] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to!, 2023. URL: https://arxiv.org/abs/2310.03693, arXiv:2310.03693.

[301] Gauthama Raman M. R., Chuadhry Mujeeb Ahmed, and Aditya Mathur. Machine learning for intrusion detection in industrial control systems: Challenges and lessons from experimental evaluation. *Cybersecurity*, 4(27), 2021. URL: https://arxiv.org/abs/2202.11917, doi:10.48550/arXiv.2202.11917.

[302] Aida Rahmattalabi, Shahin Jabbari, Himabindu Lakkaraju, Phebe Vayanos, Max Izenberg, Ryan Brown, Eric Rice, and Milind Tambe. Fair influence maximization: A welfare optimization approach. In *Proceedings of the AAAI Conference on Artificial Intelligence 35th*, 2021. URL: https://arxiv.org/abs/2006.07906, doi:10.48550/arXiv.2006.07906.

[303] Adnan Siraj Rakin, Md Hafizul Islam Chowdhuryy, Fan Yao, and Deliang Fan. DeepSteal: Advanced model extractions leveraging efficient weight stealing in memories. In *2022 IEEE Symposium on Security and Privacy (S&P)*, pages 1157–1174, 2022. URL: https://ieeexplore.ieee.org/document/9833743, doi:10.1109/SP46214.2022.9833743.

[304] Dhanesh Ramachandram and Graham W. Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34(6):96–108, 2017. URL: https://ieeexplore.ieee.org/document/8103116, doi:10.1109/MSP.2017.2738401.

[305] Javier Rando and Florian Tramèr. Universal jailbreak backdoors from poisoned human feedback, 2024. URL: https://arxiv.org/abs/2311.14455, arXiv:2311.14455, doi:10.48550/arXiv.2311.14455.

[306] Traian Rebedea, Razvan Dinu, Makesh Sreedhar, Christopher Parisien, and Jonathan Cohen. NeMo Guardrails: A toolkit for controllable and safe LLM applications with programmable rails, 2023. URL: https://arxiv.org/abs/2310.10501, arXiv:2310.10501, doi:10.48550/arXiv.2310.10501.

和 Sewoong Oh。释放随机化在审计差分隐私机器学习中的力量。载于神经信息处理系统进展,2023年。URL: https: //arxiv.org/abs/2305.18447,doi:10.48550/arXiv.2305.18447.

[298] PromptArmor 和 Kai Greshake。通过间接提示注入从 writer.com 进行数据窃取，2023。URL:https://PromptArmor.substack.com/p/data-exfil tration-from-writercom。

[299] 乔纳森·普罗岑科、布莱恩·帕诺、艾默里克·弗罗姆赫兹、克里斯·霍布利策尔、玛丽娜·波尔-ubelova、Karthikeyan Bhargavan、Benjamin Beurdouche、Joonwon Choi、Antoine Delignat-Lavaud、Cédric Fournet、Natalia Kulatova、Tahina Ramananandro、Aseem Rastogi、Nikhil Swamy、Christoph Wintersteiger 和 Santiago Zanella-Beguelin。EverCrypt：一个快速、经过验证的跨平台密码学提供者。载于*IEEE* 安全与隐私研讨会*(Oakland)*论文集,2020年5月。URL: https: //eprint.iacr.org/2019/757,doi:10.1109/SP40000.2020.00114.

[300] Xiangyu Qi、Yi Zeng、Tinghao Xie、Pin-Yu Chen、Ruoxi Jia、Prateek Mittal 和 Peter Henderson。对齐语言模型的微调会损害安全性，即使用户并无此意图!,2023。URL:https://arxiv.org/abs/2310.03693, arXiv:2310.03693。

[301] Gauthama Raman M.R.、Chuadhry Mujeeb Ahmed 和 Aditya Mathur。工业控制系统入侵检测的机器学习：实验评估的挑战与经验教训。*Cybersecurity*,4(27),2021。URL:https://arxiv.org/abs/2202.11917，doi：10.48550/arXiv.2202.11917。

[302] 艾达·拉赫马塔拉比、沙欣·贾巴里、希马宾杜·拉卡卡朱、菲比·瓦亚诺斯、马克斯·伊森-berg、Ryan Brown、Eric Rice 和 Milind Tambe。公平影响力最大化：一种福利优化方法。载于第*35*届 *AAAI* 人工智能会议论文集,2021年。URL: https://arxiv.org/abs/2006.07906, doi: 10.48550/arXiv.2006.07906.

[303] Adnan Siraj Rakin、Md Hafizul Islam Chowdhuryy、Fan Yao 和 Deliang Fan。DeepSteal：利用内存中高效权重窃取的先进模型萃取。载于*2022*年 *IEEE* 安全与隐私研讨会*(S&P)*，第1157-1174页，2022年。URL: https://ieeexplore.ieee.org/document/9833743, doi:10.1109/SP46214.2022.9833743.

[304] Dhanesh Ramachandram 和 Graham W.Taylor。深度多模态学习：一项综述近期进展与趋势综述。*IEEE* 信号处理杂志,34(6): 96-108,2017年。URL: https://ieeexplore.ieee.org/document/8103116, doi:10.1109/MSP.2017.2738401。

[305] Javier Rando 和 Florian Tramèr。来自被投毒的人类数据的通用越狱后门男性反馈，2024年。网址：https://arxiv.org/abs/2311.14455, arXiv:2311.14455, doi:10.48550/arXiv.2311.14455。

[306] Traian Rebedea、Razvan Dinu、Makesh Sreedhar、Christopher Parisien 和 Jonathan Cohen。NeMo Guardrails：一个用于构建具有可编程护栏的可控且安全 LLM 应用的工具包，2023。URL:https://arxiv.org/abs/2310.10501, arXiv:2310.1 0501,doi:10.48550/arXiv.2310.10501.

[307] Johann Rehberger. Data exfiltration via markdown injection - exploiting chatgpt's webpilot plugin, May 16 2023. Embrace The Red, Accessed: 2024-08-18. URL: https://embracethered.com/blog/posts/2023/chatgpt-webpilot-data-exfil-via-markdown-injection/.

[308] SNYK Report. AI Code Security and Trust: Organizations must change their approach. https://go.snyk.io/2023-ai-code-security-report-dwn-typ.html?aliId=eyJpIjoiUDFvdzRSdHI0dm5rVktvSSIsInQiOiJxOElRU2dQdkdqQm03ZjNLSDFFVkxBPT0ifQ%253D%253D, 2023. Human Centered AI, Stanford University.

[309] Maria Rigaki and Sebastian Garcia. A survey of privacy attacks in machine learning. *ACM Comput. Surv.*, 56(4), November 2023. doi:10.1145/3624010.

[310] Maria Rigaki and Sebastian Garcia. A survey of privacy attacks in machine learning. *ACM Comput. Surv.*, 56(4), November 2023. doi:10.1145/3624010.

[311] Rishi Bommasani, et al. On the opportunities and risks of foundation models, 2024. URL: https://arxiv.org/abs/2108.07258, arXiv:2108.07258.

[312] Alexander Robey, Eric Wong, Hamed Hassani, and George J Pappas. SmoothLLM: Defending large language models against jailbreaking attacks. *ArXiv*, abs/2310.03684, 2023. URL: https://api.semanticscholar.org/CorpusID:263671542, doi:10.48550/arXiv.2310.03684.

[313] Robust Intelligence. AI Firewall, 2023. URL: https://www.robustintelligence.com/platform/ai-firewall.

[314] Elan Rosenfeld, Ezra Winston, Pradeep Ravikumar, and Zico Kolter. Certified robustness to label-flipping attacks via randomized smoothing. In *International Conference on Machine Learning*, pages 8230–8241. PMLR, 2020. URL: https://arxiv.org/abs/1902.02918, doi:10.48550/arXiv.1902.02918.

[315] Benjamin IP Rubinstein, Blaine Nelson, Ling Huang, Anthony D Joseph, Shing-hon Lau, Satish Rao, Nina Taft, and J Doug Tygar. Antidote: understanding and defending against poisoning of anomaly detectors. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement*, pages 1–14, 2009. doi:10.1145/1644893.1644895.

[316] Mark Russinovich, Ahmed Salem, and Ronen Eldan. Great, now write an article about that: The crescendo multi-turn LLM jailbreak attack. *ArXiv*, abs/2404.01833, 2024. URL: https://api.semanticscholar.org/CorpusID:268856920, doi:10.48550/arXiv.2404.01833.

[317] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Hervé Jégou. White-box vs black-box: Bayes optimal strategies for membership inference. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 5558–5567. PMLR, 2019. URL: https://arxiv.org/abs/1908.11229, doi:10.48550/arXiv.1908.11229.

[318] Carl Sabottke, Octavian Suciu, and Tudor Dumitras. Vulnerability disclosure in the age of social media: Exploiting Twitter for predicting real-world exploits. In *24th USENIX Security Symposium (USENIX Security 15)*, pages 1041–1056, Washington, D.C., August 2015. USENIX Association. URL: https://www.usenix.org/conference/

[307] Johann Rehberger。通过 Markdown 注入进行数据外泄--利用 ChatGPT 的 WebPilot 插件，2023 年 5 月 16 日。Embrace The Red，访问日期：2024-08-18。URL:https://Embrace The Red.com/blog/posts/2023/chatgpt-webpilot-data-exfil-via-m arkdown-injection/。

[308] SNYK 报告。AI 代码安全与信任：组织必须改变其方法。https://go.snyk.io/2023-ai-code-security-report-dwn-typ.html?aliId=eyJpIjoiUDFvdzRSdHI0dm5rVktvSSIsInQiOiJxOElRU2dQdkdqQm03ZjNLSDFFVkxBPT0ifQ%253D%253D,2023。以人为本的人工智能，斯坦福大学。

[309] Maria Rigaki 和 Sebastian Garcia。机器学习中的隐私攻击综述。*ACM Comput.Surv.*,56(4),2023年11月。doi:10.1145/3624010.

[310] Maria Rigaki 和 Sebastian Garcia。机器学习中的隐私攻击综述。*ACM Comput.Surv.*,56(4),2023年11月。doi:10.1145/3624010.

[311] Rishi Bommasani 等，《论基础模型的机遇与风险》，2024 年。网址：https://arxiv.org/abs/2108.07258, arXiv：2108.07258。

[312] Alexander Robey、Eric Wong、Hamed Hassani和George J Pappas。SmoothLLM：Defend防御大型语言模型免受越狱攻击。*ArXiv*,abs/2310.03684,2023。网址：https://api.semanticscholar.org/CorpusID:263671542, doi:10.48550/arXiv.2310.03684.

[313] Robust Intelligence。AI 防火墙，2023 年。网址：https://www.Robust Intelligence.com/p平台/ai-firewall。

[314] Elan Rosenfeld、Ezra Winston、Pradeep Ravikumar 和 Zico Kolter。经认证的鲁棒-通过随机平滑对标签翻转攻击的鲁棒性。载于国际机器学习会议，第8230-8241页。PMLR,2020年。URL: https://arxiv.org/abs/ 1902.02918,doi:10.48550/arXiv.1902.02918.

[315] Benjamin IP Rubinstein、Blaine Nelson、Ling Huang、Anthony D Joseph、Shing-hon Lau、Satish Rao、Nina Taft 和 J Doug Tygar。Antidote：理解并防御异常检测器的投毒。载于第*9*届 *ACM SIGCOMM* 互联网测量会议论文集，第1-14页，2009年。doi:10.1145/1644893.1644895。

[316] Mark Russinovich、Ahmed Salem 和 Ronen Eldan。很好，现在写一篇关于这个的文章：Crescendo 多轮 LLM 越狱攻击。*ArXiv*,abs/2404.01833,2024。URL:https://api.semanticscholar.org/CorpusID:268856920, doi:10.48550 /arXiv.2404.01833。

[317] Alexandre Sablayrolles、Matthijs Douze、Cordelia Schmid、Yann Ollivier 和 Hervé Jégou。白盒 vs 黑盒：成员推断的贝叶斯最优策略。载于*ICML*，第97卷，机器学习研究论文集，第5558-5567页。PMLR,2019年。URL: https://arxiv.org/abs/1908.11229, doi:10.48550/arXiv.190 8.11229。

[318] Carl Sabottke、Octavian Suciu 和 Tudor Dumitras。社交媒体时代的漏洞披露：利用 Twitter 预测现实世界的漏洞利用。载于第*24*届 *USENIX* 安全研讨会*(USENIX Security 15)*，第1041-1056页，华盛顿特区，2015年8月。USENIX 协会。URL: https://www.usenix.org/conference/

usenixsecurity15/technical-sessions/presentation/sabottke.

[319] Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can ai-generated text be reliably detected?, 2024. URL: https://arxiv.org/abs/2303.11156, arXiv:2303.11156, doi:10.48550/arXiv.2303.11156.

[320] Vinu Sankar Sadasivan, Shoumik Saha, Gaurang Sriramanan, Priyatham Kattakinda, Atoosa Chegini, and Soheil Feizi. Fast adversarial attacks on language models in one gpu minute, 2024. URL: https://arxiv.org/abs/2402.15570, arXiv:2402.15570, doi:10.48550/arXiv.2402.15570.

[321] Ahmed Salem, Giovanni Cherubin, David Evans, Boris Köpf, Andrew Paverd, Anshuman Suri, Shruti Tople, and Santiago Zanella-Béguelin. SoK: Let the privacy games begin! A unified treatment of data inference privacy in machine learning. https://arxiv.org/abs/2212.10986, 2022. doi:10.48550/ARXIV.2212.10986.

[322] Ahmed Salem, Rui Wen, Michael Backes, Shiqing Ma, and Yang Zhang. Dynamic backdoor attacks against machine learning models. https://arxiv.org/abs/2003.03675, 2020. doi:10.48550/ARXIV.2003.03675.

[323] Roman Samoilenko. New prompt injection attack on ChatGPT web version. markdown images can steal your chat data, 2023. URL: https://systemweakness.com/new-prompt-injection-attack-on-chatgpt-web-version-ef717492c5c2.

[324] Scale AI. Adversarial robustness leaderboard. https://scale.com/leaderboard/adversarial_robustness, 2024. Accessed: 2024-08-22.

[325] Oscar Schwartz. In 2016, Microsoft's racist chatbot revealed the dangers of online conversation: The bot learned language from people on Twitter—but it also learned values. https://spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation, 2019. IEEE Spectrum.

[326] R. Schwartz, A. Vassilev, K. Greene, L. Perine, A. Burt, and P. Hall. Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. https://doi.org/10.6028/NIST.SP.1270, 2022. Special Publication (NIST SP) 800-1270, National Institute of Standards and Technology, Gaitherburg, MD. URL: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf, doi:10.6028/NIST.SP.1270.

[327] Avi Schwarzschild, Micah Goldblum, Arjun Gupta, John P Dickerson, and Tom Goldstein. Just how toxic is data poisoning? A unified benchmark for backdoor and data poisoning attacks. https://arxiv.org/abs/2006.12557, 2020. arXiv. doi:10.48550/ARXIV.2006.12557.

[328] Leo Schwinn, David Dobre, Sophie Xhonneux, Gauthier Gidel, and Stephan Gunnemann. Soft prompt threats: Attacking safety alignment and unlearning in open-source LLMs through the embedding space, 2024. URL: https://arxiv.org/abs/2402.09063, arXiv:2402.09063, doi:10.48550/arXiv.2402.09063.

[329] Giorgio Severi, Jim Meyer, Scott Coull, and Alina Oprea. Explanation-guided backdoor poisoning attacks against malware classifiers. In *30th USENIX Security Symposium (USENIX Security 2021)*, 2021. URL: https://www.usenix.org/conference/usenixsecurity21/presentation/severi.

[330] Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor

usenixsecurity15/技术会议/演示/sabottke。

[319] 维努·桑卡尔·萨达西万、奥农·库马尔、斯里拉姆·巴拉苏布拉马尼安、王文晓，和 Soheil Feizi。ai 生成的文本能被可靠检测吗?,2024年。URL: https://arxi v.org/abs/2303.11156,arXiv:2303.11156,doi:10.48550/arXiv.2303.11156.

[320] [222] Vinu Sankar Sadasivan、Shoumik Saha、Gaurang Sriramanan、Priyatham Kattakinda、Atoosa Chegini 和 Soheil Feizi。在一个 GPU 分钟内对语言模型进行快速对抗攻击，2024 年。URL: arXiv:2402.15570,

doi:10.48550/arXiv.2402.15570。

[321] 艾哈迈德·萨勒姆，乔瓦尼·凯鲁宾，大卫·埃文斯，鲍里斯·科普夫，安德鲁·帕弗德，安-Shuman Suri、Shruti Tople 和 Santiago Zanella-Béguelin。SoK：让隐私博弈开始!机器学习中数据推断隐私的统一处理。https://arxiv.org/abs/2212.10986, 2022。doi:10.48550/ARXIV.2212.10986.

[322] Ahmed Salem、Rui Wen、Michael Backes、Shiqing Ma 和 Yang Zhang。针对机器学习模型的动态后门攻击。https://arxiv.org/abs/2003.036 75,2020。doi:10.48550/ARXIV.2003.03675.

[323] Roman Samoilenko.针对ChatGPT网页版的新型提示注入攻击。markdown图片可窃取你的聊天数据，2023。URL: https://systemweakness.com/n ew-prompt-injection-attack-on-chatgpt-web-version-ef717492c5c2。

[324] Scale AI。对抗鲁棒性排行榜。https://scale.com/leaderboard/adve 对抗鲁棒性，2024。访问日期：2024-08-22。

[325] Oscar Schwartz.2016年，微软的种族主义聊天机器人揭示了在线对话的危险：该机器人从Twitter上的人那里学习语言--但它也学习了价值观。https://spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the - dangers-of-online-conversation,2019。IEEE Spectrum。

[326] R. Schwartz、A. Vassilev、K. Greene、L. Perine、A. Burt和P. Hall。迈向识别和管理人工智能偏见的标准。https://doi.org/10.602 8/NIST.SP.1270，2022年。特别出版物（NIST SP）800-1270，美国国家标准与技术研究院，马里兰州盖瑟斯堡。网址：https://nvlpubs.nist.gov/nistp ubs/SpecialPublications/NIST.SP.1270.pdf，doi：10.6028/NIST.SP.1270.

[327] Avi Schwarzschild、Micah Goldblum、Arjun Gupta、John P Dickerson和Tom Goldstein.数据投毒究竟有多毒?后门与数据投毒攻击的统一基准。https://arxiv.org/abs/2006.12557, 2020。arXiv。

doi:10.48550/ARXIV.2006.12557。

[328] Leo Schwinn、David Dobre、Sophie Xhonneux、Gauthier Gidel和Stephan Gunnemann.软提示威胁：通过嵌入空间攻击开源LLM中的安全对齐与遗忘，2024。URL:

https://arxiv.org/abs/2402. 09063,arXiv:2402.09063,doi:10.48550/arXiv.2402.09063.

[329] Giorgio Severi、Jim Meyer、Scott Coull 和 Alina Oprea。针对恶意软件分类器的解释引导后门投毒攻击。载于第*30*届 *USENIX* 安全研讨会*(USENIX Security 2021)*,2021年。URL: https://www.usenix.org/conference/us ENIX Security21/presentation/severi。

[330] 阿里·沙法希，黄荣尼，马赫亚尔·纳吉比，奥克塔维安·苏丘，克里斯托夫·斯图德，图多尔

Dumitras, and Tom Goldstein. Poison frogs! Targeted clean-label poisoning attacks on neural networks. In *Advances in Neural Information Processing Systems*, pages 6103–6113, 2018. URL: https://arxiv.org/abs/1804.00792, doi:10.48550/arXiv.1804.00792.

[331] Shawn Shan, Arjun Nitin Bhagoji, Haitao Zheng, and Ben Y. Zhao. Poison forensics: Traceback of data poisoning attacks in neural networks. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 3575–3592, Boston, MA, August 2022. USENIX Association. URL: https://www.usenix.org/conference/usenixsecurity22/presentation/shan.

[332] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 23rd ACM SIGSAC Conference on Computer and Communications Security*, October 2016. URL: https://www.ece.cmu.edu/~lbauer/papers/2016/ccs2016-face-recognition.pdf, doi:10.1145/2976749.2978392.

[333] Vasu Sharma, Ankita Kalra, Vaibhav, Simral Chaudhary, Labhesh Patel, and LP Morency. Attend and attack: Attention guided adversarial attacks on visual question answering models. https://nips2018vigil.github.io/static/papers/accepted/33.pdf, 2018.

[334] Ryan Sheatsley, Blaine Hoak, Eric Pauley, Yohan Beugin, Michael J. Weisman, and Patrick McDaniel. On the robustness of domain constraints. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, CCS '21, page 495–515, New York, NY, USA, 2021. Association for Computing Machinery. doi:10.1145/3460120.3484570.

[335] Virat Shejwalkar and Amir Houmansadr. Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning. In *NDSS*, 2021. URL: https://www.ndss-symposium.org/wp-content/uploads/ndss2021_6C-3_24498_paper.pdf.

[336] Virat Shejwalkar, Amir Houmansadr, Peter Kairouz, and Daniel Ramage. Back to the drawing board: A critical evaluation of poisoning attacks on production federated learning. In *43rd IEEE Symposium on Security and Privacy, SP 2022, San Francisco, CA, USA, May 22-26, 2022*, pages 1354–1371. IEEE, 2022. doi:10.1109/SP46214.2022.9833647.

[337] Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. " do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. *arXiv preprint arXiv:2308.03825*, 2023. URL: https://arxiv.org/abs/2308.03825, doi:10.48550/arXiv.2308.03825.

[338] Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. *CoRR*, abs/2308.03825, 2023. URL: https://doi.org/10.48550/arXiv.2308.03825, arXiv:2308.03825, doi:10.48550/ARXIV.2308.03825.

[339] Xinyue Shen, Yiting Qu, Michael Backes, and Yang Zhang. Prompt stealing attacks against text-to-image generation models. *arXiv preprint arXiv:2302.09923*, 2023.

Dumitras 和 Tom Goldstein。毒蛙!针对神经网络的定向干净标签投毒攻击。载于神经信息处理系统进展，第6103-6113页，2018年。URL: https://arxiv.org/abs/1804.00792, doi:10.48550/arXiv.1804.00792。

[331] Shawn Shan、Arjun Nitin Bhagoji、Haitao Zheng 和 Ben Y.Zhao。投毒取证：神经网络中数据投毒攻击的追溯。载于第*31*届 *USENIX* 安全研讨会*(USENIX Security 22)*，第3575-3592页，波士顿，马萨诸塞州，2022年8月。USENIX 协会。URL: https://www.usenix.org/conference/USENIX Security22/presentat ion/shan。

[332] Mahmood Sharif、Sruti Bhagavatula、Lujo Bauer 和 Michael K.Reiter。为犯罪配饰：对最先进人脸识别的真实且隐蔽的攻击。载于第*23*届 *ACM SIGSAC* 计算机与通信安全会议论文集,2016年10月。URL: https://www.ece.cmu.edu/~lbauer/papers/2016/cc s2016-face-recognition.pdf,doi:10.1145/2976749.2978392.

[333] Vasu Sharma, Ankita Kalra, Vaibhav, Simral Chaudhary, Labhesh Patel, and LP Morency.关注并攻击：针对视觉问答模型的注意力引导对抗攻击。https://nips2018vigil.github.io/static/papers/accepted/3 3.pdf,2018。

[334] Ryan Sheatsley、Blaine Hoak、Eric Pauley、Yohan Beugin、Michael J.Weisman 和 Patrick McDaniel。论域约束的鲁棒性。载于*2021*年 *ACM SIGSAC* 计算机与通信安全会议论文集,CCS '21，第495-515页，纽约，纽约州，美国，2021年。美国计算机协会。doi:10.1145/3460120.3484570.

[335] Virat Shejwalkar and Amir Houmansadr.操纵拜占庭：优化联邦学习的模型投毒攻击与防御。In *NDSS*,2021。URL: https://www.ndss-symposium.org/wp-content/uploads/ndss2021_6C-3_24498_p aper.pdf。

[336] Virat Shejwalkar、Amir Houmansadr、Peter Kairouz 和 Daniel Ramage。回到绘图板：对生产联邦学习中投毒攻击的批判性评估。载于第*43*届 *IEEE* 安全与隐私研讨会，*SP 2022*，旧金山，加利福尼亚州，美国，*2022*年*5*月*22-26*日，第1354-1371页。IEEE,2022年。doi:10.1109/SP4621 4.2022.9833647。

[337] 沈新月、陈泽源、Michael Backes、沈云、张洋。“do anything now”：刻画和评估大语言模型上的真实世界越狱提示。*arXiv* 预印本 *arXiv:2308.03825*,2023年。URL: https://arxiv.org/ab

[338] 沈新月、陈泽源、Michael Backes、沈云、张阳。“do anything now”：刻画与评估大语言模型上的真实世界越狱提示。*CoRR*,abs/2308.03825,2023。URL: https://doi.org/10.48550/arXiv.2308.03825,arXiv:2308.03825,doi:10.48550/ARXIV.2308.03825.

[339] Xinyue Shen、Yiting Qu、Michael Backes 和 Yang Zhang。针对文本到图像生成模型的提示窃取攻击。*arXiv* 预印本 *arXiv:2302.09923*,2023年。

URL: https://arxiv.org/abs/2302.09923, doi:10.48550/arXiv.2302.09923.

[340] Abhay Sheshadri, Aidan Ewart, Phillip Guo, Aengus Lynch, Cindy Wu, Vivek Hebbar, Henry Sleight, Asa Cooper Stickland, Ethan Perez, Dylan Hadfield-Menell, and Stephen Casper. Targeted latent adversarial training improves robustness to persistent harmful behaviors in LLMs, 2024. URL: https://arxiv.org/abs/2407.15549, arXiv:2407.15549, doi:10.48550/arXiv.2407.15549.

[341] Cong Shi, Tianfang Zhang, Zhuohang Li, Huy Phan, Tianming Zhao, Yan Wang, Jian Liu, Bo Yuan, and Yingying Chen. Audio-domain position-independent backdoor attack via unnoticeable triggers. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*, MobiCom '22, page 583–595, New York, NY, USA, 2022. Association for Computing Machinery. doi:10.1145/3495243.3560531.

[342] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2017. URL: https://arxiv.org/abs/1610.05820, doi:10.48550/arXiv.1610.05820.

[343] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy (S&P), Oakland*, 2017. URL: https://arxiv.org/abs/1610.05820, doi:10.48550/arXiv.1610.05820.

[344] Satya Narayan Shukla, Anit Kumar Sahu, Devin Willmott, and Zico Kolter. Simple and efficient hard label black-box adversarial attacks in low query budget regimes. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 1461–1469, New York, NY, USA, 2021. Association for Computing Machinery. doi:10.1145/3447548.3467386.

[345] Ilia Shumailov, Yiren Zhao, Daniel Bates, Nicolas Papernot, Robert Mullins, and Ross Anderson. Sponge examples: Energy-latency attacks on neural networks. https://arxiv.org/abs/2006.03463, 2020. doi:10.48550/ARXIV.2006.03463.

[346] Gagandeep Singh, Timon Gehr, Markus Püschel, and Martin Vechev. An abstract domain for certifying neural networks. *Proc. ACM Program. Lang.*, 3, January 2019. doi:10.1145/3290354.

[347] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. URL: https://arxiv.org/abs/2001.07685, doi:10.48550/arXiv.2001.07685.

[348] Saleh Soltan, Shankar Ananthakrishnan, Jack FitzGerald, Rahul Gupta, Wael Hamza, Haidar Khan, Charith Peris, Stephen Rawls, Andy Rosenbaum, Anna Rumshisky, Chandana Satya Prakash, Mukund Sridhar, Fabian Triefenbach, Apurv Verma, Gokhan Tur, and Prem Natarajan. AlexaTM 20B: Few-shot learning using a large-scale multilingual seq2seq model. https://www.amazon.science/publications/al

URL: https://arxiv.org/abs/2302.09923, doi:10.48550/arXiv.2302.09923.

[340] Abhay Sheshadri, Aidan Ewart, Phillip Guo, Aengus Lynch, Cindy Wu, Vivek Heb- bar, Henry Sleight, Asa Cooper Stickland, Ethan Perez, Dylan Hadfield-Menell, and Stephen Casper.定向潜在对抗训练提升LLM对持续有害行为的鲁棒性，2024。URL: https://arxiv.org/abs/2407.15549, arXiv:2407.15549,doi:10.48550/arXiv.2407.15549.

[341] Cong Shi、Tianfang Zhang、Zhuohang Li、Huy Phan、Tianming Zhao、Yan Wang、Jian Liu、Bo Yuan 和 Yingying Chen。通过不可察觉触发器的音频域位置无关后门攻击。载于第*28*届移动计算与网络国际会议论文集,MobiCom '22，第583-595页，纽约，纽约州，美国，2022年。美国计算机协会。doi:10.1145/3495243.3560531.

[342] Reza Shokri、Marco Stronati、Congzheng Song 和 Vitaly Shmatikov。针对机器学习模型的成员推断攻击。载于*2017*年 *IEEE* 安全与隐私研讨会*(SP)*，第3-18页。IEEE,2017年。URL: https://arxiv.org/abs/1610. 05820,doi:10.48550/arXiv.1610.05820.

[343] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov.针对机器学习模型的成员推断攻击。In *IEEE Symposium on Security and Privacy (S&P) , Oakland*, 2017. URL: https://arxiv.org/abs/1610.05820, doi: 10.48550/arXiv.1610.05820.

[344] Satya Narayan Shukla, Anit Kumar Sahu, Devin Willmott, and Zico Kolter.低查询预算机制下简单高效的硬标签黑盒对抗攻击.In 第 *27* 届 *ACM SIGKDD* 知识发现与数据挖掘会议论文集, KDD '21， 第 1461-1469 页， 美国纽约州纽约市， 2021. 美国计算机协会.doi:10.1145/3447548.3467386.

[345] Ilia Shumailov, Yiren Zhao, Daniel Bates, Nicolas Papernot, Robert Mullins, and Ross Anderson.海绵样本：针对神经网络的能耗-延迟攻击。https: //arxiv.org/abs/2006.03463,2020。doi:10.48550/ARXIV.2006.03463.

[346] Gagandeep Singh, Timon Gehr, Markus Püschel, and Martin Vechev.用于认证神经网络的抽象域。*Proc.ACM Program.Lang.*, 3, January 2019. doi:10.1145/3290354。

[347] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. FixMatch：利用一致性和置信度简化半监督学习。载于《第34届国际神经信息处理系统会议论文集》，NIPS'20，美国纽约州红钩，2020年。Curran Associates Inc. URL: https://arxiv.org/abs/2001.07685, doi:10.48550/arXiv.2001.07685。

[348] Saleh Soltan, Shankar Ananthakrishnan, Jack FitzGerald, Rahul Gupta, Wael Hamza, Haidar Khan, Charith Peris, Stephen Rawls, Andy Rosenbaum, Anna Rumshisky, Chandana Satya Prakash, Mukund Sridhar, Fabian Triefenbach, Apurv Verma, Gokhan Tur, and Prem Natarajan. AlexaTM 20B：使用大规模多语言seq2seq模型的小样本学习。https://www.amazon.science/publications/al

exatm-20b-few-shot-learning-using-a-large-scale-multilingual-seq2seq-model, 2022. Amazon.

[349] Dawn Song, Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Florian Tramèr, Atul Prakash, and Tadayoshi Kohno. Physical adversarial examples for object detectors. In *12th USENIX Workshop on Offensive Technologies (WOOT 18)*, Baltimore, MD, August 2018. USENIX Association. URL: https://www.usenix.org /conference/woot18/presentation/eykholt, doi:10.48550/arXiv.1807.07769.

[350] Shuang Song and David Marn. Introducing a new privacy testing library in TensorFlow, 2020. URL: https://blog.tensorflow.org/2020/06/introducing-new-privacy-testing-library.html.

[351] Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. A strongreject for empty jailbreaks, 2024. URL: https://arxiv.org/abs/2402.10260, arXiv:2402.10260, doi:10.48550/arXiv.2402.10260.

[352] N. Srndic and P. Laskov. Practical evasion of a learning-based classifier: A case study. In *Proc. IEEE Security and Privacy Symposium*, 2014. URL: https://personal.utdallas.edu/~muratk/courses/dmsec_files/srndic-laskov-sp2014.pdf.

[353] U.S. AI Safety Institute Technical Staff. Strengthening ai agent hijacking evaluations, 2024. URL: https://www.nist.gov/news-events/news/2025/01/technical-blog-strengthening-ai-agent-hijacking-evaluations.

[354] Jacob Steinhardt, Pang Wei W Koh, and Percy S Liang. Certified defenses for data poisoning attacks. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/9d7311ba459f9e45ed746755a32dcd11-Paper.pdf, doi:10.48550/arXiv.1706.03691.

[355] Thomas Steinke, Milad Nasr, and Matthew Jagielski. Privacy auditing with one (1) training run. In *Advances in Neural Information Processing Systems*, 2023. URL: https://arxiv.org/abs/2305.08846, doi:10.48550/arXiv.2305.08846.

[356] Ellen Su, Anu Vellore, Amy Chang, Raffaele Mura, Blaine Nelson, Paul Kassianik, and Amin Karbasi. Extracting memorized training data via decomposition. *arXiv preprint arXiv:2409.12367*, 2024. URL: https://arxiv.org/abs/2409.12367, doi:10.48550/arXiv.2409.12367.

[357] Octavian Suciu, Scott E Coull, and Jeffrey Johns. Exploring adversarial examples in malware detection. In *2019 IEEE Security and Privacy Workshops (SPW)*, pages 8–14. IEEE, 2019. URL: https://arxiv.org/abs/1810.08280, doi:10.48550/arXiv.1810.08280.

[358] Octavian Suciu, Radu Marginean, Yigitcan Kaya, Hal Daume III, and Tudor Dumitras. When does machine learning FAIL? generalized transferability for evasion and poisoning attacks. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 1299–1316, 2018. URL: https://arxiv.org/abs/1803.06975, doi:10.48550/arXiv.1803.06975.

exatm-20b-使用大规模多语言seq2seq模型的少样本学习，2022. 亚马逊。

[349] Dawn Song, Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Florian Tramèr, Atul Prakash，和 Tadayoshi Kohno。面向目标检测器的物理对抗样本。载于第12届USENIX攻击技术研讨会（WOOT 18），马里兰州巴尔的摩，2018年8月。USENIX协会。URL：https://www.usenix.org /conference/woot18/presentation/eykholt，doi：10.48550/arXiv.1807.07769.

[350] 宋爽和大卫·马恩。在TensorFlow中引入新的隐私测试库，2020年。网址：https://blog.TensorFlow.org/2020/06/Introducing-new-privacy-t esting-library.html。

[351] Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. 针对空洞越狱的强拒绝，2024年。URL: https://arxiv.org/abs/2402.10260, arXiv:2402.10260, doi:10.48550/arXiv.2402.10260.

[352] N.Srndic and P.Laskov.对基于学习的分类器的实际规避：案例研究。In *Proc.IEEE Security and Privacy Symposium*, 2014. URL: https://personal.utdalla s.edu/~muratk/courses/dmsec_files/srndic-laskov-sp2014.pdf.

[353] 美国人工智能安全研究所技术人员。加强 AI 智能体劫持评估，2024 年。网址：https://www.nist.gov/news-events/news/2025/01/technical-blog-str engthening-ai-agent-hijacking-evaluations。

[354] Jacob Steinhardt, Pang Wei W Koh, and Percy S Liang. 针对数据投毒攻击的认证防御。载于I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett编，《神经信息处理系统进展》，第30卷。Curran Associates, Inc.，2017年。URL: https://proceedings.ne urips.cc/paper/2017/file/9d7311ba459f9e45ed746755a32dcd11-Paper.pdf，doi:10.48550/arXiv.1706.03691。

[355] Thomas Steinke, Milad Nasr, and Matthew Jagielski.单次训练运行的隐私审计。In *Advances in Neural Information Processing Systems*, 2023. URL: https://arxiv.org/abs/2305.08846, doi:10.48550/arXiv.2305.08846.

[356] Ellen Su, Anu Vellore, Amy Chang, Raffaele Mura, Blaine Nelson, Paul Kassianik, and Amin Karbasi.通过分解提取记忆的训练数据。*arXiv preprint arXiv:2409.12367*, 2024. URL: https://arxiv.org/abs/2409.12367, doi:10.48550 /arXiv.2409.12367.

[357] Octavian Suciu, Scott E Coull, and Jeffrey Johns.探索恶意软件检测中的对抗样本。In *2019 IEEE Security and Privacy Workshops (SPW)*, pages 8-14. IEEE, 2019. URL: https://arxiv.org/abs/1810.08280, doi:10.48550/arXiv.1810. 08280.

[358] Octavian Suciu, Radu Marginean, Yigitcan Kaya, Hal Daume III, and Tudor Dumitras. 机器学习何时失效?规避与投毒攻击的泛化可迁移性。In *27th USENIX Security Symposium (USENIX Security 18)*, pages 1299-1316, 2018. URL: https://arxiv.org/abs/1803.06975, doi:10.48550/arXiv.1803.06975.

[359] Jingwei Sun, Ang Li, Louis DiValentin, Amin Hassanzadeh, Yiran Chen, and Hai Li. FL-WBC: Enhancing robustness against model poisoning attacks in federated learning from a client perspective. In *NeurIPS*, 2021. URL: https://arxiv.org/abs/2110.13864, doi:10.48550/arXiv.2110.13864.

[360] Ziteng Sun, Peter Kairouz, Ananda Theertha Suresh, and H Brendan McMahan. Can you really backdoor federated learning? *arXiv:1911.07963*, 2019. URL: https://arxiv.org/abs/1911.07963, doi:10.48550/arXiv.1911.07963.

[361] Anshuman Suri and David Evans. Formalizing and estimating distribution inference risks. *Proceedings on Privacy Enhancing Technologies*, 2022. URL: https://arxiv.org/abs/2109.06024, doi:10.48550/arXiv.2109.06024.

[362] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. URL: http://arxiv.org/abs/1312.6199, doi:10.48550/arXiv.1312.6199.

[363] Rahim Taheri, Reza Javidan, Mohammad Shojafar, Zahra Pooranian, Ali Miri, and Mauro Conti. On defending against label flipping attacks on malware detection systems. *CoRR*, abs/1908.04473, 2019. URL: http://arxiv.org/abs/1908.04473, arXiv:1908.04473, doi:10.48550/arXiv.1908.04473.

[364] Azure AI Red Team. Pyrit: The python risk identification tool for generative ai. https://github.com/Azure/PyRIT, 2024. Accessed: 2024-08-18.

[365] The Llama Team. The LLaMA3 Herd of Models, 2024. URL: https://arxiv.org/abs/2407.21783, doi:10.48550/arXiv.2407.21783.

[366] The White House. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/, October 2023. The White House.

[367] T. Ben Thompson and Michael Sklar. Breaking circuit breakers. URL: https://confirmlabs.org/posts/circuit_breaking.html.

[368] T. Ben Thompson and Michael Sklar. Fluent student-teacher redteaming, 2024. URL: https://arxiv.org/abs/2407.17447, arXiv:2407.17447, doi:10.48550/arXiv.2407.17447.

[369] Anvith Thudi, Ilia Shumailov, Franziska Boenisch, and Nicolas Papernot. Bounding membership inference. https://arxiv.org/abs/2202.12232, 2022. doi:10.48550/ARXIV.2202.12232.

[370] Lionel Nganyewou Tidjon and Foutse Khomh. Threat assessment in machine learning based systems. *arXiv preprint arXiv:2207.00091*, 2022. URL: https://arxiv.org/abs/2207.00091, doi:10.48550/arXiv.2207.00091.

[371] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models, 2023. URL: https://arxiv.org/abs/2302

[359] 孙经纬、李昂、Louis DiValentin、Amin Hassanzadeh、陈一然和Hai Li。FL-WBC：从客户端视角增强联邦学习中对模型投毒攻击的鲁棒性。In *NeurIPS*, 2021. URL: https://arxiv.org/abs/2110.13864,

[360] Ziteng Sun、Peter Kairouz、Ananda Theertha Suresh 和 H Brendan McMahan。你真的能在联邦学习中植入后门吗?*arXiv:1911.07963*,2019 年。URL:https://arxi v.org/abs/1911.07963,doi:10.48550/arXiv.1911.07963.

[361] Anshuman Suri 和 David Evans。形式化与估计分布推断风险。*Proceedings on Privacy Enhancing Technologies*,2022。URL:https://arxiv.or g/abs/2109.06024,doi:10.48550/arXiv.2109.06024.

[362] Christian Szegedy、Wojciech Zaremba、Ilya Sutskever、Joan Bruna、Dumitru Erhan、Ian Goodfellow 和 Rob Fergus。神经网络的引人注目的特性。载于国际学习表征会议,2014年。网址：http://arxiv.org/abs/ 1312.6199,doi:10.48550/arXiv.1312.6199.

[363] Rahim Taheri, Reza Javidan, Mohammad Shojafar, Zahra Pooranian, Ali Miri, and Mauro Conti.论防御恶意软件检测系统上的标签翻转攻击。*CoRR*, abs/1908.04473, 2019. URL: http://arxiv.org/abs/1908.04473, arXiv:1908.04473, doi:10.48550/arXiv.1908.04473.

[364] Azure AI 红队。Pyrit：生成式 AI 的 Python 风险识别工具。https://github.com/Azure/PyRIT,2024。访问日期：2024-08-18。

[365] Llama 团队。LLaMA3 模型群，2024 年。网址：https://arxiv.org/abs/2407.21783 ，doi:10.48550/arXiv.2407.21783。

[366] 白宫。关于安全、可靠和可信开发--的行政命令人工智能的开发与使用。https://www.White House.gov/briefing-roo m/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-t rustworthy-development-and-use-of-Artificial-Intelligence/, 2023年10月。白宫。

[367] T.Ben Thompson 和 Michael Sklar。攻破断路器。URL: https://confirmlabs.org/posts/circuit_breaking.html。

[368] T.Ben Thompson 和 Michael Sklar。流畅的学生-教师红队测试，2024。URL: https://arxiv.org/abs/2407.17447, arXiv:2407.17447, doi:10.48550/arXiv.2 407.17447。

[369] Anvith Thudi, Ilia Shumailov, Franziska Boenisch, and Nicolas Papernot. 限定成员推断。https://arxiv.org/abs/2202.12232, 2022年。doi:10.48550 /ARXIV.2202.12232。

[370] Lionel Nganyewou Tidjon 和 Foutse Khomh。机器学习中的威胁评估--基于.....的系统。*arXiv*预印本 *arXiv:2207.00091*,2022年。网址：https://arxiv.org/

[371] Hugo Touvron、Thibaut Lavril、Gautier Izacard、Xavier Martinet、Marie-Anne Lachaux、Timothée Lacroix、Baptiste Rozière、Naman Goyal、Eric Hambro、Faisal Azhar、Aurelien Rodriguez、Armand Joulin、Edouard Grave 和 Guillaume Lample。LLaMA：开放且高效的基础语言模型，2023年。网址：https://arxiv.org/abs/2302

.13971, arXiv:2302.13971, doi:10.48550/arXiv.2302.13971.

[372] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL: https://arxiv.org/abs/2307.09288, arXiv:2307.09288, doi:10.48550/arXiv.2307.09288.

[373] Florian Tramer. Detecting adversarial examples is (Nearly) as hard as classifying them. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 21692–21702. PMLR, 17–23 Jul 2022. URL: https://proceedings.mlr.press/v162/tramer22a.html.

[374] Florian Tramer, Jens Behrmann, Nicholas Carlini, Nicolas Papernot, and Joern-Henrik Jacobsen. Fundamental tradeoffs between invariance and sensitivity to adversarial perturbations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9561–9571. PMLR, 13–18 Jul 2020. URL: https://proceedings.mlr.press/v119/tramer20a.html, doi:10.48550/arXiv.2002.04599.

[375] Florian Tramèr, Nicholas Carlini, Wieland Brendel, and Aleksander Mądry. On adaptive attacks to adversarial example defenses. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. URL: https://arxiv.org/abs/2002.08347, doi:10.48550/arXiv.2002.08347.

[376] Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction APIs. In *USENIX Security*, 2016. URL: https://arxiv.org/abs/1609.02943, doi:10.48550/arXiv.1609.02943.

[377] Florian Tramèr, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. The space of transferable adversarial examples. https://arxiv.org/abs/1704.03453, 2017. doi:10.48550/ARXIV.1704.03453.

[378] Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31.

.13971, arXiv:2302.13971, doi:10.48550/arXiv.2302.13971。

[372] Hugo Touvron、Louis Martin、Kevin Stone、Peter Albert、Amjad Almahairi、Yasmine Babaei、Nikolay Bashlykov、Soumya Batra、Prajjwal Bhargava、Shruti Bhosale、Dan Bikel、Lukas Blecher、Cristian Canton Ferrer、Moya Chen、Guillem Cucurull、David Esiobu、Jude Fernandes、Jeremy Fu、Wenyin Fu、Brian Fuller、Cynthia Gao、Vedanuj Goswami、Naman Goyal、Anthony Hartshorn、Saghar Hosseini、Rui Hou、Hakan Inan、Marcin Kardas、Viktor Kerkez、Madian Khabsa、Isabel Kloumann、Artem Korenev、Punit Singh Koura、Marie-Anne Lachaux、Thibaut Lavril、Jenya Lee、Diana Liskovich、Yinghai Lu、Yuning Mao、Xavier Martinet、Todor Mihaylov、Pushkar Mishra、Igor Molybog、Yixin Nie、Andrew Poulton、Jeremy Reizenstein、Rashi Rungta、Kalyan Saladi、Alan Schelten、Ruan Silva、Eric Michael Smith、Ranjan Subramanian、Xiaoqing Ellen Tan、Binh Tang、Ross Taylor、Adina Williams、Jian Xiang Kuan、Puxin Xu、Zheng Yan、Iliyan Zarov、Yuchen Zhang、Angela Fan、Melanie Kambadur、Sharan Narang、Aurelien Rodriguez、Robert Stojnic、Sergey Edunov 和 Thomas Scialom。Llama 2：开放基础模型与微调聊天模型，2023年。网址：https://arxiv.org/abs/2307.09288, arXiv:2307.09288，doi:10.48550/arXiv.2307.09288.

[373] Florian Tramer。检测对抗样本几乎与对其进行分类一样困难。载于 Kamalika Chaudhuri、Stefanie Jegelka、Le Song、Csaba Szepesvari、Gang Niu 和 Sivan Sabato 编，第*39*届国际机器学习会议论文集，第162卷，机器学习研究论文集，第21692-21702页。PMLR,2022年7月17-23日。URL: https://Proceedings.mlr.press/v162/tramer22a.html。

[374] Florian Tramer、Jens Behrmann、Nicholas Carlini、Nicolas Papernot 和 Joern-Henrik Jacobsen。不变性与对抗扰动敏感性之间的基本权衡。载于 Hal Daumé III 和 Aarti Singh 编，第*37*届国际机器学习会议论文集，第119卷，机器学习研究会议论文集，第9561-9571页。PMLR,2020年7月13-18日。网址：https://pr oceedings.mlr.press/v119/tramer20a.html,doi:10.48550/arXiv.2002.04599.

[375] Florian Tramèr、Nicholas Carlini、Wieland Brendel 和 Aleksander Mądry。论自适应针对对抗样本防御的主动攻击。载于第*34*届国际神经信息处理系统会议论文集,NIPS'20,Red Hook,NY,USA,2020。Curran Associates Inc.URL: https://arxiv.org/abs/2002.08347,

[376] 弗洛里安·特拉梅尔、张帆、阿里·朱尔斯、迈克尔·K·赖特和托马斯·里斯滕帕特。窃取-通过预测 API 攻击机器学习模型。载于*USENIX Security*,2016 年。网址：https://arxiv.org/abs/1609.02943, doi:10.48550/arXiv.1609.02943.

[377] Florian Tramèr, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel.可迁移对抗样本的空间.https://arxiv.org/abs/1704.03453, 2017. doi:10.48550/ARXIV.1704.03453.

[378] 布兰登·特兰、杰瑞·李和亚历山大·马德里。后门攻击中的谱特征--攻击。载于 S.Bengio、H.Wallach、H.Larochelle、K.Grauman、N.Cesa-Bianchi 和 R.Garnett 编，*Advances in Neural Information Processing Systems*，第 31 卷。

Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper/2018/file/280cf18baf4311c92aa5a042336587d3-Paper.pdf, doi:10.48550/arXiv.1811.00636.

[379] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019. URL: https://openreview.net/forum?id=SyxAb30cY7, doi:10.48550/arXiv.1805.12152.

[380] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Clean-label backdoor attacks. In *ICLR*, 2019. URL: https://openreview.net/forum?id=HJg6e2CcK7.

[381] U.K. AI Safety Institute. Advanced ai evaluations: May update, 2024. Accessed: 2024-08-18. URL: https://www.aisi.gov.uk/work/advanced-ai-evaluations-may-update.

[382] Kush R. Varshney. *Trustworthy Machine Learning*. Independently Published, Chappaqua, NY, USA, 2022. URL: https://www.trustworthymachinelearning.com/.

[383] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. URL: https://arxiv.org/abs/1706.03762, doi:10.48550/arXiv.1706.03762.

[384] Sridhar Venkatesan, Harshvardhan Sikka, Rauf Izmailov, Ritu Chadha, Alina Oprea, and Michael J. De Lucia. Poisoning attacks and data sanitization mitigations for machine learning models in network intrusion detection systems. In *MILCOM*, pages 874–879. IEEE, 2021. URL: https://ieeexplore.ieee.org/document/9652916, doi:10.1109/MILCOM52596.2021.9652916.

[385] Sameer Wagh, Shruti Tople, Fabrice Benhamouda, Eyal Kushilevitz, Prateek Mittal, and Tal Rabin. FALCON: honest-majority maliciously secure framework for private deep learning. In *Proceedings on Privacy Enhancing Technologies (PoPETs) 2021, Issue 1*, 20201.

[386] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for attacking and analyzing NLP. *arXiv preprint arXiv:1908.07125*, 2019. URL: https://arxiv.org/abs/1908.07125, doi:10.48550/arXiv.1908.07125.

[387] Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. The instruction hierarchy: Training LLMs to prioritize privileged instructions, 2024. URL: https://arxiv.org/abs/2404.13208, arXiv:2404.13208, doi:10.48550/arXiv.2404.13208.

[388] Eric Wallace, Tony Z. Zhao, Shi Feng, and Sameer Singh. Concealed data poisoning attacks on NLP models. In *NAACL*, 2021. URL: https://arxiv.org/abs/2010.12563, doi:10.48550/arXiv.2010.12563.

[389] Alexander Wan, Eric Wallace, Sheng Shen, and Dan Klein. Poisoning language models during instruction tuning, 2023. URL: https://arxiv.org/abs/2305.00944, arXiv:2305.00944, doi:10.48550/arXiv.2305.00944.

[390] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng,

Curran Associates, Inc.，2018年。URL: https://proceedings.neurips.cc/paper/2018/file/280cf18baf4311c92aa5a042336587d3-Paper.pdf, doi:10.48550/arXiv.181 1.00636。

[379] Dimitris Tsipras、Shibani Santurkar、Logan Engstrom、Alexander Turner和Alek-桑德·马德里。鲁棒性可能与准确性相冲突。载于国际学习表征会议,2019年。网址：https://openreview.net/forum?i

[380] Alexander Turner、Dimitris Tsipras 和 Aleksander Madry。干净标签后门攻击。载于*ICLR*,2019 年。网址：https://openreview.net/forum?id=HJg6e2CcK7.

[381] 英国人工智能安全研究所。高级人工智能评估：2024年5月更新。访问日期：2024-08-18。网址：https://www.aisi.gov.uk/work/advanced-ai-evaluations-may-u pdate。

[382] Kush R.Varshney.可信机器学习.独立出版，第paqua，纽约，美国，2022年。网址：https://www.trustworthymachinelearning.com/.

[383] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 注意力就是你所需要的一切。载于《神经信息处理系统进展》，第5998-6008页，2017年。URL: https://arxiv. org/abs/1706.03762, doi:10.48550/arXiv.1706.03762.

[384] Sridhar Venkatesan、Harshvardhan Sikka、Rauf Izmailov、Ritu Chadha、Alina Oprea 和 Michael J.De Lucia。针对网络入侵检测系统中机器学习模型的投毒(poisoning)攻击与数据净化缓解措施。载于*MILCOM*，第 874-879 页。IEEE,2021 年。URL:https://ieeexplore.ieee.org/document/9652916, doi:10.1109/MILCOM52596.2021.9652916.

[385] Sameer Wagh, Shruti Tople, Fabrice Benhamouda, Eyal Kushilevitz, Prateek Mittal, and Tal Rabin.FALCON：面向私有深度学习的诚实多数恶意安全框架。载于隐私增强技术会议论文集*(PoPETs)2021*年第*1*期,20201。

[386] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh.用于攻击和分析NLP的通用对抗触发器。*arXiv*预印本*arXiv:1908.07125*,2019。URL: https://arxiv.org/abs/1908.07125, doi:10.48550/arXiv.1908.0712

5.

[387] Eric Wallace、Kai Xiao、Reimar Leike、Lilian Weng、Johannes Heidecke 和 Alex Beutel。指令层级：训练 LLM 优先处理特权指令，2024。网址：https://arxiv.org/abs/2404.13208, arXiv:2404.13208,

doi:10.48550/arXiv.2404.13208。

[388] Eric Wallace、Tony Z.Zhao、Shi Feng 和 Sameer Singh。针对 NLP 模型的隐蔽数据投毒攻击。载于*NAACL*,2021 年。URL:https://arxiv.org/abs/2010.12563, doi:10.48550/arXiv.2010.12563。

[389] @[389]@ Alexander Wan、Eric Wallace、Sheng Shen 和 Dan Klein。语言模型投毒在指令微调期间，2023年。网址：https://arxiv.org/abs/2305.00944, arXiv:2305.00944,doi:10.48550/arXiv.2305.00944.

[390] 王博伦、姚远顺、单珊、李慧颖、Bimal Viswanath、郑海涛，

and Ben Y. Zhao. Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 707–723, San Francisco, CA, USA, May 2019. IEEE. URL: https://ieeexplore.ieee.org/document/8835365/, doi:10.1109/SP.2019.00031.

[391] Haotao Wang, Tianlong Chen, Shupeng Gui, Ting-Kuei Hu, Ji Liu, and Zhangyang Wang. Once-for-All Adversarial Training: In-Situ Tradeoff between Robustness and Accuracy for Free. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada*, 2020. URL: https://arxiv.org/abs/2010.11828, doi:10.48550/arXiv.2010.11828.

[392] Hongyi Wang, Kartik Sreenivasan, Shashank Rajput, Harit Vishwakarma, Saurabh Agarwal, Jy-yong Sohn, Kangwook Lee, and Dimitris Papailiopoulos. Attack of the Tails: Yes, You Really Can Backdoor Federated Learning. In *NeurIPS*, 2020. URL: https://arxiv.org/abs/2007.05084, doi:10.48550/arXiv.2007.05084.

[393] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6), March 2024. URL: http://dx.doi.org/10.1007/s11704-024-40231-1, doi:10.1007/s11704-024-40231-1.

[394] Shiqi Wang, Kexin Pei, Justin Whitehouse, Junfeng Yang, and Suman Jana. Formal security analysis of neural networks using symbolic intervals. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 1599–1614, Baltimore, MD, August 2018. USENIX Association. URL: https://www.usenix.org/conference/usenixsecurity18/presentation/wang-shiqi.

[395] Wenxiao Wang, Alexander Levine, and Soheil Feizi. Improved certified defenses against data poisoning with (deterministic) finite aggregation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 22769–22783. PMLR, 2022. URL: https://proceedings.mlr.press/v162/wang22m.html, doi:10.48550/arXiv.2202.02628.

[396] Wenxiao Wang, Alexander J Levine, and Soheil Feizi. Improved certified defenses against data poisoning with (Deterministic) finite aggregation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 22769–22783. PMLR, 17–23 Jul 2022. URL: https://proceedings.mlr.press/v162/wang22m.html, doi:10.48550/arXiv.2202.02628.

[397] Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 1924–1933. Computer Vision Foundation / IEEE, 2021. URL: https://openaccess.thecvf.com/content/CVPR2021/html/Wang_Enhancing_the_Transferability_of_Adversarial_Attacks_Through_Va

以及 Ben Y.Zhao。Neural Cleanse：识别与缓解神经网络中的后门攻击。载于*2019 IEEE Symposium on Security and Privacy (SP)*，第 707-723 页，San Francisco, CA, USA,2019 年 5 月。IEEE。URL: https://ieeexplore.ieee.org/document/8835365/，doi:10.1109/SP.2019.00031。

[391] 王浩涛、陈天龙、桂书鹏、胡廷魁、刘骥、王张扬。《一劳永逸的对抗训练：稳健性与准确性的免费原位权衡》。载于第34届神经信息处理系统大会（NeurIPS 2020）论文集，加拿大温哥华，2020年。网址：https://arxiv.org/abs/2010.11828，doi:10.48550/arXiv.2010.11828.

[392] Hongyi Wang、Kartik Sreenivasan、Shashank Rajput、Harit Vishwakarma、Saurabh Agarwal、Jy-yong Sohn、Kangwook Lee 和 Dimitris Papailiopoulos。尾部攻击：是的，你确实可以在联邦学习中植入后门。载于*NeurIPS*,2020年。URL:https://arxiv.org/abs/2007.05084, doi:10.48550/arXiv.2007.05084.

[393] 王磊、马晨、冯雪阳、张泽宇、杨浩、张景森、陈志远、唐佳凯、陈旭、林彦凯、赵鑫、魏哲伟和文继荣。基于大语言模型的自主智能体综述。《计算机科学前沿》，18(6)，2024年3月。网址：http://dx.doi.org/10.1007/s11704-024-40231-1, doi:10.1007/s11704-024-40231-1.

[394] 王世奇、裴可欣、Justin Whitehouse、杨俊峰和Suman Jana。使用符号区间对神经网络进行形式化安全分析。载于第27届USENIX安全研讨会（USENIX Security 18），第1599-1614页，马里兰州巴尔的摩，2018年8月。USENIX协会。网址：https://www.usenix.org/conference/usenixsecuri ty18/presentation/wang-shiqi。

[395] 王文晓、Alexander Levine 和 Soheil Feizi。利用（确定性）有限聚合改进针对数据投毒的认证防御。载于 Kamalika Chaudhuri、Stefanie Jegelka、Le Song、Csaba Szepesvári、Gang Niu 和 Sivan Sabato 编，《国际机器学习会议，ICML 2022，2022年7月17-23日，美国马里兰州巴尔的摩》，第162卷《机器学习研究论文集》，第22769-22783页。PMLR，2022年。网址：https://proceedings.mlr.press/ v162/wang22m.html，doi:10.48550/arXiv.2202.02628.

[396] 王文晓、Alexander J Levine 和 Soheil Feizi。利用（确定性）有限聚合改进针对数据投毒的认证防御。载于 Kamalika Chaudhuri、Stefanie Jegelka、Le Song、Csaba Szepesvari、Gang Niu 和 Sivan Sabato 编，《第39届国际机器学习会议论文集》，《机器学习研究论文集》第162卷，第22769-22783页。PMLR，2022年7月17-23日。网址：https://proceedings.mlr.press/v162/wang22m.html, doi:10.48550/arXiv.2202.02628。

[397] 王小森和何鲲。通过方差调优增强对抗攻击的可迁移性。载于IEEE计算机视觉与模式识别会议，CVPR 2021，虚拟会议，2021年6月19-25日，第1924-1933页。计算机视觉基金会 / IEEE，2021年。网址：https://openaccess.thecvf.com/content/CVPR2021 /html/Wang_Enhancing_the_Transferability_of_Adversarial_Attacks_Through_Va

riance_Tuning_CVPR_2021_paper.html, doi:10.1109/CVPR46437.2021.00196.

[398] Yanting Wang, Wei Zou, and Jinyuan Jia. FCert: Certifiably robust few-shot classification in the era of foundation models. In *Proc. IEEE Security and Privacy Symposium*, 2024. URL: https://arxiv.org/abs/2404.08631, doi:10.48550/arXiv.2404.08631.

[399] Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. Do-not-answer: A dataset for evaluating safeguards in LLMs, 2023. URL: https://arxiv.org/abs/2308.13387, arXiv:2308.13387, doi:10.48550/arXiv.2308.13387.

[400] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does LLM safety training fail? *arXiv preprint arXiv:2307.02483*, 2023.

[401] Xingxing Wei, Jun Zhu, Sha Yuan, and Hang Su. Sparse adversarial perturbations for videos. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019. doi:10.1609/aaai.v33i01.33018973.

[402] Zhipeng Wei, Jingjing Chen, Xingxing Wei, Linxi Jiang, Tat-Seng Chua, Fengfeng Zhou, and Yu-Gang Jiang. Heuristic black-box adversarial attacks on video recognition models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12338–12345, 2020. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6918, doi:10.48550/arXiv.1911.09449.

[403] Lilian Weng. Adversarial attacks on latent language models, 2023. URL: https://lilianweng.github.io/posts/2023-10-25-adv-attack-llm/.

[404] Emily Wenger, Josephine Passananti, Arjun Nitin Bhagoji, Yuanshun Yao, Haitao Zheng, and Ben Y. Zhao. Backdoor attacks against deep learning systems in the physical world. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6202–6211, 2020. URL: https://arxiv.org/abs/2006.14580, doi:10.48550/arXiv.2006.14580.

[405] Simon Willison. The dual LLM pattern for building AI assistants that can resist prompt injection, 2023. Accessed: 2024-08-22. URL: https://simonwillison.net/2023/Apr/25/dual-llm-pattern/.

[406] Yotam Wolf, Noam Wies, Yoav Levine, and Amnon Shashua. Fundamental limitations of alignment in large language models. *ArXiv*, abs/2304.11082, 2023. URL: https://api.semanticscholar.org/CorpusID:258291526, doi:10.48550/arXiv.2304.11082.

[407] Dongxian Wu and Yisen Wang. Adversarial neuron pruning purifies backdoored deep models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 16913–16925. Curran Associates, Inc., 2021. URL: https://proceedings.neurips.cc/paper/2021/file/8cbe9ce23f42628c98f80fa0fac8b19a-Paper.pdf, doi:10.48550/arXiv.2110.14430.

[408] Fangzhou Wu, Ning Zhang, Somesh Jha, Patrick McDaniel, and Chaowei Xiao. A new

方差调优_CVPR_2021_论文.html，doi:10.1109/CVPR46437.2021.00196.

[398] Yanting Wang、Wei Zou 和 Jinyuan Jia。FCert：可认证鲁棒的少样本分类基础模型时代的 tion。载于*Proc.IEEE Security and Privacy Symposium*,2024。URL:https://arxiv.org/abs/2404.08631, doi:10.48550/arXiv.2404.08631.

[399] 王玉霞、李浩楠、韩旭东、普雷斯拉夫·纳科夫和蒂莫西·鲍德温。请勿-answer：用于评估 LLM 防护措施的数据集，2023年。URL:https://arxiv.org/abs/2308.13387,arXiv:2308.13387,doi:10.48550/arXiv.2308.13387.

[400] Alexander Wei、Nika Haghtalab 和 Jacob Steinhardt。Jailbroken：LLM 安全训练为何失败？arXiv 预印本 arXiv:2307.02483，2023 年。

[401] 魏星星、朱军、袁莎、苏航。视频的稀疏对抗扰动。载于《第三十三届AAAI人工智能会议、第三十一届人工智能创新应用会议及第九届AAAI教育人工智能进展研讨会论文集》，AAAI'19/IAAI'19/EAAI'19。AAAI出版社，2019年。doi:10.1609/aaai.v33i01.33018973.

[402] 魏志鹏、陈晶晶、魏星星、蒋林希、蔡达成、周峰峰和蒋玉刚。针对视频识别模型的启发式黑盒对抗攻击。载于《AAAI人工智能会议论文集》，第34卷，第12338-12345页，2020年。网址：https://ojs.aaai.org/index.php/AAAI/article/view /6918，doi:10.48550/arXiv.1911.09449.

[403] Lilian Weng。针对潜在语言模型的对抗攻击，2023年。URL:https://lilianweng.github.io/posts/2023-10-25-adv-attack-llm/.

[404] Emily Wenger、Josephine Passananti、Arjun Nitin Bhagoji、Yuanshun Yao、Haitao Zheng 和 Ben Y. Zhao。针对物理世界中深度学习系统的后门攻击。2021 IEEE/CVF 计算机视觉与模式识别会议（CVPR），第 6202-6211 页，2020 年。网址：https://arxiv.org/abs/2006.14580, doi:10.48550/arXiv.2006.14580.

[405] Simon Willison。用于构建可抵御提示注入的 AI 助手的双 LLM 模式，2023年。访问日期：2024-08-22。URL:https://Simon Willison.net/2023/Apr /25/dual-llm-pattern/。

[406] Yotam Wolf、Noam Wies、Yoav Levine 和 Amnon Shashua。大语言模型中对齐的根本局限性。*ArXiv*,abs/2304.11082,2023年。URL:https://api.semanticscholar.org/CorpusID:258291526,doi:10.48550/arXiv.2304.11 082。

[407] 吴东贤和王一森。对抗性神经元剪枝净化后门深度模型。载于 M. Ranzato、A. Beygelzimer、Y. Dauphin、P.S. Liang 和 J. Wortman Vaughan 编，《神经信息处理系统进展》，第 34 卷，第 16913-16925 页。Curran Associates, Inc.，2021 年。网址：https://proceedings. neurips.cc/paper/2021/file/8cbe9ce23f42628c98f80fa0fac8b19a-Paper.pdf，doi:10.48550/arXiv.2110.14430.

[408] Fangzhou Wu、Ning Zhang、Somesh Jha、Patrick McDaniel 和 Chaowei Xiao。LLM 安全的新

era in LLM security: Exploring security concerns in real-world LLM-based systems, 2024. URL: https://arxiv.org/abs/2402.18649, arXiv:2402.18649.

[409] Xi Wu, Matthew Fredrikson, Somesh Jha, and Jeffrey F. Naughton. A methodology for formalizing model-inversion attacks. In *2016 IEEE 29th Computer Security Foundations Symposium (CSF)*, pages 355–370, 2016. doi:10.1109/CSF.2016.32.

[410] Yuhao Wu, Franziska Roesner, Tadayoshi Kohno, Ning Zhang, and Umar Iqbal. SecGPT: An execution isolation architecture for LLM-based systems, 2024. URL: https://arxiv.org/abs/2403.04960, arXiv:2403.04960, doi:10.48550/arXiv.2402.18649.

[411] Zhen Xiang, David J. Miller, and George Kesidis. Post-training detection of backdoor attacks for two-class and multi-attack scenarios. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL: https://openreview.net/forum?id=MSgB8D4Hy51, doi:10.48550/arXiv.2201.08474.

[412] Huang Xiao, Battista Biggio, Gavin Brown, Giorgio Fumera, Claudia Eckert, and Fabio Roli. Is feature selection secure against training data poisoning? In *International Conference on Machine Learning*, pages 1689–1698, 2015. URL: https://arxiv.org/abs/1804.07933, doi:10.48550/arXiv.1804.07933.

[413] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6256–6268. Curran Associates, Inc., 2020. URL: https://proceedings.neurips.cc/paper/2020/file/44feb0096faa8326192570788b38c1d1-Paper.pdf, doi:10.48550/arXiv.1904.12848.

[414] Weilin Xu, Yanjun Qi, and David Evans. Automatically evading classifiers. In *Proceedings of the 2016 Network and Distributed Systems Symposium*, pages 21–24, 2016. URL: https://www.cs.virginia.edu/~evans/pubs/ndss2016/.

[415] Xiaojun Xu, Xinyun Chen, Chang Liu, Anna Rohrbach, Trevor Darrell, and Dawn Song. Fooling vision and language models despite localization and attention mechanism. https://arxiv.org/abs/1709.08693, 2017. doi:10.48550/ARXIV.1709.08693.

[416] Xiaojun Xu, Qi Wang, Huichen Li, Nikita Borisov, Carl A. Gunter, and Bo Li. Detecting AI trojans using meta neural analysis. In *IEEE Symposium on Security and Privacy, S&P 2021*, pages 103–120, United States, May 2021. URL: https://ieeexplore.ieee.org/document/9519467, doi:10.1109/SP40001.2021.00034.

[417] Karren Yang, Wan-Yi Lin, Manash Barman, Filipe Condessa, and Zico Kolter. Defending multimodal fusion models against single-source adversaries. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Xplore, 2022. URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9578130, doi:10.48550/ARXIV.2206.12714.

[418] Limin Yang, Zhi Chen, Jacopo Cortellazzi, Feargus Pendlebury, Kevin Tu, Fabio Pierazzi, Lorenzo Cavallaro, and Gang Wang. Jigsaw puzzle: Selective backdoor attack to subvert malware classifiers. *CoRR*, abs/2202.05470, 2022. URL: https://arxiv.or

时代：探索现实世界中基于 LLM 的系统的安全问题，2024. 网址：https://arxiv.org/abs/2402.18649, arXiv:2402.18649。

[409] 吴曦、Matthew Fredrikson、Somesh Jha 和 Jeffrey F. Naughton。形式化模型反演攻击的方法论。载于 2016 年 IEEE 第 29 届计算机安全基础研讨会（CSF），第 355-370 页，2016 年。doi:10.1109/CSF.2016.32.

[410] Yuhao Wu、Franziska Roesner、Tadayoshi Kohno、Ning Zhang 和 Umar Iqbal。SecGPT：一种面向基于 LLM 的系统的执行隔离架构，2024年。URL:https://arxiv.org/abs/2403.04960, arXiv:2403.04960,doi:10.48550/arX iv.2402.18649。

[411] 甄翔、David J. Miller 和 George Kesidis。两类和多攻击场景下后门攻击的训练后检测。载于第十届国际学习表征会议，ICLR 2022，虚拟活动，2022年4月25-29日。OpenReview.net，2022。网址：https://openreview.net/forum?id=MSgB8D4Hy51, doi:10.48550/arXiv.2201.08474.

[412] 黄晓、Battista Biggio、Gavin Brown、Giorgio Fumera、Claudia Eckert 和 Fabio Roli。特征选择对训练数据投毒是否安全？载于《国际机器学习会议论文集》，第1689-1698页，2015年。网址：https://arxiv.org/abs/1804.07933,doi:10.48550/arXiv.1804.07933.

[413] 谢启哲、戴子航、Eduard Hovy、Thang Luong 和 Quoc Le。用于一致性训练的无监督数据增强。载于 H. Larochelle、M. Ranzato、R. Hadsell、M.F. Balcan 和 H. Lin 编，《神经信息处理系统进展》，第 33 卷，第 6256-6268 页。Curran Associates, Inc.，2020 年。网址：https://proceeding s.neurips.cc/paper/2020/file/44feb0096faa8326192570788b38c1d1-Paper.pdf，doi:10.48550/arXiv.1904.12848.

[414] Weilin Xu, Yanjun Qi, and David Evans.自动规避分类器。载于*Proceed-*[391]2016年网络与分布式系统研讨会论文集，第21-24页，2016年。网址：https://www.cs.virginia.edu/~evans/pubs/ndss2016/.

[415] Xiaojun Xu、Xinyun Chen、Chang Liu、Anna Rohrbach、Trevor Darrell 和 Dawn Song。在存在定位与注意力机制的情况下欺骗视觉与语言模型。https://arxiv.org/abs/1709.08693, 2017年。doi:10.48550/ARXIV.1709.08693.

[416] 徐晓军、王琦、李慧晨、Nikita Borisov、Carl A. Gunter 和 李波。使用元神经分析检测人工智能木马。载于 IEEE 安全与隐私研讨会，S&P 2021，第 103-120 页，美国，2021 年 5 月。网址：https://ieeexplore.ieee. org/document/9519467，doi:10.1109/SP40001.2021.00034.

[417] 卡伦·杨、林万怡、马纳什·巴曼、菲利佩·孔德萨和齐科·科尔特。Defend-针对单源对手的多模态融合模型。载于*2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*。IEEE Xplore,2022。URL: https://IEEE Xplore.ieee.org/stamp/stamp.jsp?arnumber=9578130,

[418] 杨丽敏、陈志、雅各布·科尔特拉齐、费尔格斯·彭德尔伯里、凯文·图、法比奥·皮耶尔-azzi、Lorenzo Cavallaro 和 Gang Wang。拼图：选择性后门攻击以颠覆恶意软件分类器。*CoRR*,abs/2202.05470,2022。URL:https://arxiv.or

g/abs/2202.05470, arXiv:2202.05470, doi:10.48550/arXiv.2202.05470.

[419] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models, 2023. URL: https://arxiv.org/abs/2210.03629, arXiv:2210.03629, doi:10.48550/arXiv.2210.03629.

[420] Yuanshun Yao, Huiying Li, Haitao Zheng, and Ben Y. Zhao. Latent backdoor attacks on deep neural networks. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, CCS '19, page 2041–2055, New York, NY, USA, 2019. Association for Computing Machinery. doi:10.1145/3319535.3354209.

[421] Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. Enhanced membership inference attacks against machine learning models. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, CCS '22, page 3093–3106, New York, NY, USA, 2022. Association for Computing Machinery. doi:10.1145/3548606.3560675.

[422] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *IEEE Computer Security Foundations Symposium*, CSF '18, pages 268–282, 2018. https://arxiv.org/abs/1709.01604. URL: https://arxiv.org/abs/1709.01604, doi:10.48550/arXiv.1709.01604.

[423] Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates. In *ICML*, 2018. URL: https://arxiv.org/abs/1803.01498, doi:10.48550/arXiv.1803.01498.

[424] Youngjoon Yu, Hong Joo Lee, Byeong Cheon Kim, Jung Uk Kim, and Yong Man Ro. Investigating vulnerability to adversarial examples on multimodal data fusion in deep learning. https://arxiv.org/abs/2005.10987, 2020. Online. doi:10.48550/ARXIV.2005.10987.

[425] Andrew Yuan, Alina Oprea, and Cheng Tan. Dropout attacks. In *IEEE Symposium on Security and Privacy (S&P)*, 2024. URL: https://arxiv.org/abs/2309.01614, doi:10.48550/arXiv.2309.01614.

[426] Santiago Zanella-Béguelin, Lukas Wutschitz, Shruti Tople, Victor Rühle, Andrew Paverd, Olga Ohrimenko, Boris Köpf, and Marc Brockschmidt. Analyzing information leakage of updates to natural language models. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, page 363–375, New York, NY, USA, 2020. Association for Computing Machinery. doi:10.1145/3372297.3417880.

[427] Santiago Zanella-Beguelin, Lukas Wutschitz, Shruti Tople, Ahmed Salem, Victor Rühle, Andrew Paverd, Mohammad Naseri, Boris Köpf, and Daniel Jones. Bayesian estimation of differential privacy. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 40624–40636. PMLR, 23–29 Jul

g/abs/2202.05470, arXiv:2202.05470, doi:10.48550/arXiv.2202.05470.

[419] Shunyu Yao、Jeffrey Zhao、Dian Yu、Nan Du、Izhak Shafran、Karthik Narasimhan 和 Yuan Cao。React：在语言模型中协同推理与行动，2023。URL:https://arxiv.org/abs/2210.03629, arXiv:2210.03629,doi:10.48550/arXiv.2 210.03629。

[420] 姚远顺、李慧颖、郑海涛和Ben Y. Zhao。深度神经网络上的潜在后门攻击。载于2019年ACM SIGSAC计算机与通信安全会议论文集，CCS '19，第2041-2055页，美国纽约州纽约市，2019年。美国计算机协会。doi:10.1145/3319535.3354209.

[421] Jiayuan Ye、Aadyaa Maddi、Sasi Kumar Murakonda、Vincent Bindschaedler和Reza Shokri。针对机器学习模型的增强成员推断攻击。载于2022年ACM SIGSAC计算机与通信安全会议论文集，CCS '22，第3093-3106页，美国纽约州纽约市，2022年。美国计算机协会。doi:10.1145/3548606.3560675.

[422] Samuel Yeom、Irene Giacomelli、Matt Fredrikson和Somesh Jha。机器学习中的隐私风险：分析与过拟合的关联。载于IEEE计算机安全基础研讨会，CSF '18，第268-282页，2018年。https://arxiv.org/abs/17 09.01604。URL: https://arxiv.org/abs/1709.01604, doi:10.48550/arXiv.1709.01604.

[423] Dong Yin、Yudong Chen、Ramchandran Kannan 和 Peter Bartlett。拜占庭鲁棒分布式学习：迈向最优统计速率。载于*ICML*,2018。URL:https: //arxiv.org/abs/1803.01498,doi:10.48550/arXiv.1803.01498.

[424] Youngjoon Yu、Hong Joo Lee、Byeong Cheon Kim、Jung Uk Kim 和 Yong Man Ro。探究深度学习中多模态数据融合对对抗样本的脆弱性。https://arxiv.org/abs/2005.10987, 2020。在线。doi:10.48550/ARXIV.2005.10987.

[425] Andrew Yuan、Alina Oprea和Cheng Tan。丢弃攻击。载于IEEE安全与隐私研讨会（S&P），2024年。URL: https://arxiv.org/abs/2309.01614, doi:10.48550/arXiv.2309.01614。

[426] Santiago Zanella-Béguelin、Lukas Wutschitz、Shruti Tople、Victor Rühle、Andrew Paverd、Olga Ohrimenko、Boris Köpf和Marc Brockschmidt。分析自然语言模型更新的信息泄露。载于2020年ACM SIGSAC计算机与通信安全会议论文集，第363-375页，美国纽约州纽约市，2020年。美国计算机协会。doi:10.1145/3372297.3417880.

[427] Santiago Zanella-Beguelin、Lukas Wutschitz、Shruti Tople、Ahmed Salem、Victor Rühle、Andrew Paverd、Mohammad Naseri、Boris Köpf 和 Daniel Jones。差分隐私的贝叶斯估计。载于 Andreas Krause、Emma Brunskill、Kyunghyun Cho、Barbara Engelhardt、Sivan Sabato 和 Jonathan Scarlett 编，《第40届国际机器学习会议论文集》，《机器学习研究论文集》第202卷，第40624–40636页。PMLR，7月23–29日

2023. URL: https://proceedings.mlr.press/v202/zanella-beguelin23a.html, doi:10.48550/arXiv.2206.05199.

[428] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models, 2021. URL: https://arxiv.org/abs/2106.02636, arXiv:2106.02636, doi:10.48550/arXiv.2106.02636.

[429] Yi Zeng, Si Chen, Won Park, Zhuoqing Mao, Ming Jin, and Ruoxi Jia. Adversarial unlearning of backdoors via implicit hypergradient. In *International Conference on Learning Representations*, 2022. URL: https://openreview.net/forum?id=MeeQkFYVbzW, doi:10.48550/arXiv.2110.03735.

[430] Qiusi Zhan, Zhixiang Liang, Zifan Ying, and Daniel Kang. InjecAgent: Benchmarking indirect prompt injections in tool-integrated large language model agents, 2024. URL: https://arxiv.org/abs/2403.02691, arXiv:2403.02691, doi:10.48550/arXiv.2403.02691.

[431] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115, feb 2021. doi:10.1145/3446776.

[432] Hanlin Zhang, Benjamin L. Edelman, Danilo Francati, Daniele Venturi, Giuseppe Ateniese, and Boaz Barak. Watermarks in the sand: Impossibility of strong watermarking for generative models. *ArXiv*, abs/2311.04378, 2023. URL: https://api.semanticscholar.org/CorpusID:265050535, doi:10.48550/arXiv.2311.04378.

[433] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7472–7482. PMLR, 09–15 Jun 2019. URL: https://proceedings.mlr.press/v97/zhang19p.html, doi:10.48550/arXiv.1901.08573.

[434] Ruisi Zhang, Seira Hidano, and Farinaz Koushanfar. Text revealer: Private text reconstruction via model inversion attacks against transformers. *arXiv preprint arXiv:2209.10505*, 2022. URL: https://arxiv.org/abs/2209.10505, doi:10.48550/arXiv.2209.10505.

[435] Su-Fang Zhang, Jun-Hai Zhai, Bo-Jun Xie, Yan Zhan, and Xin Wang. Multimodal representation learning: Advances, trends and challenges. In *2019 International Conference on Machine Learning and Cybernetics (ICMLC)*, pages 1–6. IEEE, 2019. URL: https://api.semanticscholar.org/CorpusID:209901378, doi:10.1109/ICMLC48188.2019.8949228.

[436] Susan Zhang, Mona Diab, and Luke Zettlemoyer. Democratizing access to large-scale language models with OPT-175B. https://ai.facebook.com/blog/democratizing-access-to-large-scale-language-models-with-opt-175b/, 2022. Meta AI.

[437] Wanrong Zhang, Shruti Tople, and Olga Ohrimenko. Leakage of dataset properties in Multi-Party machine learning. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2687–2704. USENIX Association, August 2021. URL: https://www.us

2023. URL: https://proceedings.mlr.press/v202/zanella-beguelin23a.html, doi:10.48550/arXiv.2206.05199。

[428] Rowan Zellers、Ximing Lu、Jack Hessel、Youngjae Yu、Jae Sung Park、Jize Cao、Ali Farhadi 和 Yejin Choi。Merlot：多模态神经脚本知识模型，2021。URL:https://arxiv.org/abs/2106.02636, arXiv:2106.02636,doi:10.48550/arXiv.2106.02636.

[429] Yi Zeng、Si Chen、Won Park、Zhuoqing Mao、Ming Jin和Ruoxi Jia。通过隐式超梯度对后门进行对抗性遗忘。载于国际学习表征会议，2022年。URL: https://openreview.net/forum?id=MeeQkF

YVbzW，doi:10.48550/arXiv.2110.03735.

[430] Qiusi Zhan、Zhixiang Liang、Zifan Ying 和 Daniel Kang。InjecAgent：基准-评估工具集成大语言模型智能体中的间接提示注入，2024。URL:https://arxiv.org/abs/2403.02691, arXiv:2403.02691,doi:10.48550/arX iv.2403.02691。

[431] Chiyuan Zhang、Samy Bengio、Moritz Hardt、Benjamin Recht和Oriol Vinyals。Un-解深度学习(仍然)需要重新思考泛化。*Commun.ACM*,64(3):107-115,feb 2021。doi:10.1145/3446776.

[432] 张翰林、本杰明·L·埃德尔曼、达尼洛·弗兰卡蒂、达尼埃莱·文图里、朱塞佩·阿特--niese 和 Boaz Barak。沙中之水印：生成模型强水印的不可能性。*ArXiv*,abs/2311.04378,2023。URL:https://api.semantic scholar.org/CorpusID:265050535,doi:10.48550/arXiv.2311.04378.

[433] 张宏阳、于耀东、焦剑涛、邢波、Laurent El Ghaoui 和 Michael Jordan。鲁棒性与准确性之间具有理论依据的权衡。载于 Kamalika Chaudhuri 和 Ruslan Salakhutdinov 编，《第36届国际机器学习会议论文集》，第97卷《机器学习研究论文集》，第7472-7482页。PMLR，2019年6月9日至15日。网址：https://procee dings.mlr.press/v97/zhang19p.html，doi：10.48550/arXiv.1901.08573.

[434] Ruisi Zhang、Seira Hidano和Farinaz Koushanfar。文本揭示器：通过针对变换器的模型反演攻击进行私有文本重建。arXiv预印本arXiv:2209.10505，2022年。URL: https://arxiv.org/abs/2209.10505, doi:10.485 50/arXiv.2209.10505。

[435] Su-Fang Zhang、Jun-Hai Zhai、Bo-Jun Xie、Yan Zhan 和 Xin Wang。多模态表表示学习：进展、趋势与挑战。载于*2019 International Conference on Machine Learning and Cybernetics (ICMLC)*，第 1-6 页。IEEE,2019。URL: https://api.semanticscholar.org/CorpusID:209901378, doi:10.1109/ICMLC481 88.2019.8949228。

[436] Susan Zhang、Mona Diab 和 Luke Zettlemoyer。通过 OPT-175B 实现大规模语言模型的民主化访问。https://ai.facebook.com/blog/Democratizing-acc ess-to-large-scale-language-models-with-opt-175b/,2022。Meta AI。

[437] 张婉蓉、Shruti Tople 和 Olga Ohrimenko。多参与方机器学习中数据集属性的泄露。载于第30届 USENIX 安全研讨会（USENIX Security 21），第2687-2704页。USENIX 协会，2021年8月。网址：https://www.us

enix.org/conference/usenixsecurity21/presentation/zhang-wanrong, doi:10.48550/arXiv.2006.07267.

[438] Wei Emma Zhang, Quan Z. Sheng, Ahoud Alhazmi, and Chenliang Li. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Trans. Intell. Syst. Technol.*, 11(3), apr 2020. doi:10.1145/3374217.

[439] Yiming Zhang and Daphne Ippolito. Prompts should not be seen as secrets: Systematically measuring prompt extraction attack success. *arXiv preprint arXiv:2307.06865*, 2023. URL: https://arxiv.org/abs/2307.06865, doi:10.48550/arXiv.2307.06865.

[440] Yuhao Zhang, Aws Albarghouthi, and Loris D'Antoni. Bagflip: A certified defense against data poisoning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL: https://openreview.net/forum?id=ZidkM5b92G, doi:10.48550/arXiv.2205.13634.

[441] Zhengming Zhang, Ashwinee Panda, Linyue Song, Yaoqing Yang, Michael Mahoney, Prateek Mittal, Ramchandran Kannan, and Joseph Gonzalez. Neurotoxin: Durable backdoors in federated learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 26429–26446. PMLR, 17–23 Jul 2022. URL: https://proceedings.mlr.press/v162/zhang22w.html, doi:10.48550/arXiv.2206.10341.

[442] Zhikun Zhang, Min Chen, Michael Backes, Yun Shen, and Yang Zhang. Inference attacks against graph neural networks. In *31st USENIX Security Symposium (USENIX Security 22)*, 2022. URL: https://www.usenix.org/conference/usenixsecurity22/presentation/zhang-zhikun.

[443] Junhao Zhou, Yufei Chen, Chao Shen, and Yang Zhang. Property inference attacks against GANs. In *Proceedings of Network and Distributed System Security*, NDSS, 2022. URL: https://arxiv.org/abs/2111.07608, doi:10.48550/arXiv.2111.07608.

[444] Chen Zhu, W. Ronny Huang, Hengduo Li, Gavin Taylor, Christoph Studer, and Tom Goldstein. Transferable clean-label poisoning attacks on deep neural nets. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7614–7623. PMLR, 09–15 Jun 2019. URL: https://proceedings.mlr.press/v97/zhu19a.html, doi:10.48550/arXiv.1905.05897.

[445] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences, 2020. URL: https://arxiv.org/abs/1909.08593, arXiv:1909.08593, doi:10.48550/arXiv.1909.08593.

[446] Giulio Zizzo, Chris Hankin, Sergio Maffeis, and Kevin Jones. Adversarial machine learning beyond the image domain. In *Proceedings of the 56th Annual Design Automation Conference 2019*, DAC '19, New York, NY, USA, 2019. Association for Com-

enix.org/conference/usenixsecurity21/presentation/zhang-wanrong，doi:10.48550/arXiv.2006.07267.

[438] Wei Emma Zhang、Quan Z. Sheng、Ahoud Alhazmi 和 Chenliang Li。自然语言处理中深度学习模型的对抗攻击：综述。ACM Trans. Intell. Syst. Technol.，11(3)，2020年4月。doi:10.1145/3374217.

[439] 张一鸣和达芙妮·伊波利托。提示不应被视为绝-

crets：系统性地度量提示提取攻击成功率。*arXiv preprint arXiv:2307.06865*,2023。

URL:https://arxiv.org/abs/2307.06865, doi:10.48550 /arXiv.2307.06865。

[440] Yuhao Zhang、Aws Albarghouthi 和 Loris D'Antoni。Bagflip：针对数据投毒的经认证防御。载于 Alice H.Oh、Alekh Agarwal、Danielle Belgrave 和 Kyunghyun Cho 编，*Advances in Neural Information Processing Systems*,2022。URL:https://openreview.net/forum?id=ZidkM5b92G, doi:10.48550/arXiv.220 5.13634。

[441] 张正明、Ashwinee Panda、宋林月、杨耀清、Michael Mahoney、Prateek Mittal、Ramchandran Kannan 和 Joseph Gonzalez。Neurotoxin：联邦学习中的持久后门。载于 Kamalika Chaudhuri、Stefanie Jegelka、Le Song、Csaba Szepesvari、Gang Niu 和 Sivan Sabato 编，《第39届国际机器学习会议论文集》，第162卷，《机器学习研究论文集》，第26429-26446页。PMLR，2022年7月17-23日。网址：https://proceedings.mlr.press/v162/zhang22w.html，doi：10.48550/arXiv.2206.10341.

[442] Zhikun Zhang、Min Chen、Michael Backes、Yun Shen 和 Yang Zhang。针对图神经网络的推断攻击。载于*31st USENIX Security Symposium (USENIX Security 22)*,2022。URL:https://www.usenix.org/conference/USENIX Security22/p resentation/zhang-zhikun。

[443] Junhao Zhou、Yufei Chen、Chao Shen 和 Yang Zhang。针对 GAN 的属性推断攻击。载于*Proceedings of Network and Distributed System Security*,NDSS,2022。URL:https://arxiv.org/abs/2111.07608, doi:10.48550/arXiv.2111.0760

8.

[444] Chen Zhu、W.Ronny Huang、Hengduo Li、Gavin Taylor、Christoph Studer 和 Tom Goldstein。针对深度神经网络的迁移性干净标签投毒攻击。载于 Ka- malika Chaudhuri 和 Ruslan Salakhutdinov 编，*Proceedings of the 36th In- ternational Conference on Machine Learning*，第 97 卷，*Proceedings of Ma- chine Learning Research*，第 7614-7623 页。PMLR,2019 年 6 月 09-15 日。URL:https: //Proceedings.mlr.press/v97/zhu19a.html,doi:10.48550/arXiv.1905.05897.

[445] Daniel M.Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B.Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving.基于人类偏好的语言模型微调， 2020. URL: https://arxiv.org/abs/1909.08593, arXiv:1909.08593, doi:10.48550/arXiv.1909.08593.

[446] Giulio Zizzo、Chris Hankin、Sergio Maffeis 和 Kevin Jones。超越图像域的对抗性机器学习。载于*Proceedings of the 56th Annual Design Au- tomation Conference 2019*,DAC '19,New York, NY, USA,2019。Association for Com-

puting Machinery. doi:10.1145/3316781.3323470.

[447] Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, and Dan Hendrycks. Improving alignment and robustness with circuit breakers, 2024. URL: https://arxiv.org/abs/2406.04313, arXiv:2406.04313, doi:h10.48550/arXiv.2406.04313.

[448] Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023. URL: https://arxiv.org/abs/2307.15043, doi:10.48550/arXiv.2307.15043.

[449] Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. Poisonedrag: Knowledge poisoning attacks to retrieval-augmented generation of large language models, 2024. URL: https://arxiv.org/abs/2005.11401, arXiv:2402.07867, doi:10.48550/arXiv.2005.11401.

算机协会。doi:10.1145/3316781.3323470.

[447] 安迪·邹、龙·潘、贾斯汀·王、德里克·杜埃纳斯、麦克斯韦·林、马克西姆·安-
driushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, and Dan Hendrycks.利用断路器改进对齐与鲁棒性， 2024. URL: https://arxiv.org/abs/2406.04313, arXiv:2406.04313, doi:h10.48550/arXiv.2406.04313.

[448] Andy Zou、Zifan Wang、J Zico Kolter 和 Matt Fredrikson。针对对齐语言模型的通用且可迁移的对抗攻击。*arXiv preprint arXiv:2307.15043*,2023。URL:https://arxiv.org/abs/2307.15043, doi:10.48550/arXiv.2307.1504
3.

[449] 邹伟、耿润鹏、王炳辉、贾金元。Poisonedrag：知识-
针对大语言模型检索增强生成的边缘投毒攻击， 2024. URL: https://arxiv.org/
abs/2005.11401, arXiv:2402.07867,

Appendix A. Glossary

Clicking on the page number at the end of a definition will navigate to the page where the term is used.

A

adversarial example A modified testing sample that induces misclassification or misbehavior of a machine learning model at deployment time. ix, 6

adversarial machine learning Attacks that exploit the statistical, data-based nature of machine learning systems. xii, 1

agent Software programs that can interact with their environment, receive information, and undertake self-directed actions in service of a larger, externally-specified goal. 1, 35, 37, 39, 50, 52, 54

area under the curve A measure of the ability of a classifier to distinguish between classes in machine learning. A higher AUC means that a model performs better when distinguishing between the two classes. AUC measures the entire two-dimensional area under the RECEIVER OPERATING CHARACTERISTIC (ROC) curve. 30

attribute inference attacks An attack against machine learning models that infers sensitive attributes of a training data record, given partial knowledge about the record. 7

availability breakdown In the AML context, a disruption of the ability of other users or processes to obtain timely and reliable access to an AI system's outputs or functionality. 6, 39

B

backdoor pattern A transformation or insertion applied to a data sample that triggers an adversary-specified behaviour in a model that has been subject to a backdoor poisoning attack. For example, in computer vision, an adversary could poison a model such that the insertion of a square of white pixels induces a desired target label. 6, 22, 107

backdoor poisoning attack A poisoning attack that causes a model to perform an adversary-selected behaviour in response to inputs that follow a particular BACKDOOR PATTERN. 6, 42

C

classification The task of predicting which of a set of discrete categories an input belongs to. 5

convolutional neural networks A class of feed-forward neural networks that include at least one convolutional layer, referred to as CNNs. In convolutional layers, feature detectors (known as kernels or filters) detect specific features across the input data. CNNs are primarily used for processing grid-like data, such as images, and are particularly effective for tasks like image classification, object detection, and image segmentation. 5, 31

附录 A.术语表

点击定义末尾的页码将跳转到该术语被使用的页面。

A

对抗样本一种经过修改的测试样本，可诱导错误分类或错误行为
机器学习模型在部署时的行为。ix, 6

对抗性机器学习利用机器学习基于数据的统计本质的攻击
机器学习系统。xii, 1

代理 能够与其环境交互、接收信息的软件程序，
并为服务于更大的、外部指定的目标而采取自主行动的软件程序。1, 35, 37, 39, 50, 52, 54

曲线下面积一种度量分类器区分不同类别能力的指标
AUC越高，意味着模型在区分两个类别时表现越好。AUC度量的是接收者操作特征(ROC)曲线下的整个二维面积。30

属性推断攻击一种针对机器学习模型的攻击，用于推断敏感
训练数据记录的属性，在已知该记录部分知识的情况下。7

可用性破坏在AML语境下，指其他用户或进程及时、可靠地获取人工智能系统输出或功能的能力
过程，以获取对人工智能系统输出或功能的及时可靠访问。6, 39

B

后门模式施加于数据样本的一种变换或插入，可触发已遭受后门投毒攻击的模型
中由对手指定的行为。例如，在计算机视觉中，对手可以投毒一个模型，使得插入一个白色像素方块即可诱导出期望的目标标签。6, 22, 107

后门投毒攻击一种投毒攻击，使模型执行对手
响应遵循特定后门模式的输入时所选择的行为。6, 42

C

分类预测输入属于一组离散类别中哪一类的任务
to.5

卷积神经网络 一类前馈神经网络，其包含
简称CNN。在卷积层中，特征检测器(称为核或滤波器)检测输入数据中的特定特征。CNN主要用于处理网格状数据，如图像，在图像分类、目标检测和图像分割等任务中尤为有效。5, 31

D

data confidentiality A well-established concept in cybersecurity referring to the protection of sensitive information from unauthorized access and disclosure. 7

data poisoning A POISONING ATTACKS in which an adversary controls part of the training data. 5, 36, 37, 40, 111

data privacy attacks Attacks against machine learning models that extract sensitive information about training data. 7

data reconstruction Privacy attacks that reconstruct sensitive data in a model's training data from aggregate information. 7, 28

deployment stage The stage of the machine learning pipeline in which a model is deployed into a live or real-world environment for use, such as being integrated into an enterprise application or made available to end users through an API. 5, 37, 38

diffusion models A class of latent variable generative models consisting of three major components: a forward process, a reverse process, and a sampling procedure. The goal of the diffusion model is to learn a diffusion process that generates the probability distribution of a given dataset. It is widely used in computer vision on a variety of tasks, including image denoising, inpainting, super-resolution, and image generation. 34

direct prompt injection A DIRECT PROMPTING ATTACK in which the attacker exploits PROMPT INJECTION. 43, 110

direct prompting attack In the generative AI context, an attack conducted by the primary user of the system through QUERY ACCESS (e.g., as opposed to through RESOURCE CONTROL). 34, 43, 108, 110

discriminative A type of machine learning method that learns to discriminate between classes. 5

E

energy-latency attack An attack that exploits the performance dependency on hardware and model optimizations to negate the effects of hardware optimizations, increase computational latency, increase hardware temperature, and massively increase the amount of energy consumed. 6, 8

ensemble learning A type of a meta machine learning approach that combines the predictions of several models to improve performance. 5

expectation over transformation A method for strengthening adversarial examples to remain adversarial under image transformations that occur in the real world, such as angle and viewpoint changes. EOT models these perturbations within the optimization procedure. Rather than optimizing the log-likelihood of a single example, EOT uses a chosen distribution of transformation functions that take an input controlled by the adversary to the "true" input perceived by the classifier. 16

F

D

数据机密性网络安全中一个成熟的概念，指保护
保护敏感信息免遭未经授权的访问和披露。7

数据投毒一种投毒攻击，其中对手控制部分训练
数据。5, 36, 37, 40, 111

数据隐私攻击针对机器学习模型、提取敏感信息的攻击
关于训练数据的信息。7

数据重构重构模型训练中敏感数据的隐私攻击
从汇总信息中得出的数据。7, 28

部署阶段机器学习流水线中模型被部署的阶段
投入实际运行或真实世界环境中使用，例如被集成到企业应用中或通过 API 提供给最终用户。5, 37, 38

扩散模型一类潜变量生成模型，由三个主要
组成部分构成：前向过程、反向过程和采样过程。扩散模型的目标是学习一个扩散过程，以生成给定数据集的概率分布。它广泛应用于计算机视觉领域的多种任务，包括图像去噪、图像修复、超分辨率和图像生成。34

直接提示注入一种直接提示攻击，其中攻击者利用提示
注入。43, 110

直接提示攻击在生成式人工智能语境下，由主要
用户通过查询访问(例如，相对于通过资源控制)实施的攻击。34, 43, 108, 110

判别式 一种机器学习方法，其学习区分
classes.5

E

能量-延迟攻击 一种利用性能对硬件的依赖性的攻击
和模型优化的依赖性，以抵消硬件优化的效果、增加计算延迟、提高硬件温度并大幅增加能耗量的攻击。6, 8

集成学习 一种元机器学习方法，它结合预测
几个模型的优化以提升性能。5

变换期望 一种增强对抗样本以重新
真实世界中发生的图像变换(如角度和视角变化)下仍保持对抗性。EOT 在优化过程中对这些扰动进行建模。EOT 不是优化单个样本的对数似然，而是使用选定的变换函数分布，这些函数将受攻击者控制的输入映射到分类器所感知的"真实"输入。16

F

federated learning A type of machine learning in which a model is trained in a decentralized fashion using multiple data sources without pooling or combining the data in any centralized location. Federated learning allows entities or devices to collaboratively train a global model by exchanging model updates without directly sharing the data that each entity controls. 5, 31

feedforward neural networks Artificial neural networks in which the connections between nodes is from one layer to the next and do not form a cycle. 31

fine-tuning The process of adapting a pre-trained model to perform specific tasks or specialize in a particular domain. This phase follows the initial pre-training phase and involves further training the model on task-specific data. This is often a supervised learning task. 37

fine-tuning circumvention Fine-tuning to remove model refusal behaviour or other model-level safety interventions. 41

formal methods A mathematically rigorous technique for the specification, development, and verification of software systems. 18

foundation model In generative AI, models trained on broad data using SELF-SUPERVISED LEARNING that can be adapted such as through fine-tuning for a variety of downstream tasks [311]. 111

functional attack An adversarial attack that is optimized for a set of data in a domain rather than per data point. 13, 23

G

generative adversarial networks A machine learning framework in which two neural networks contest with each other in the form of a zero-sum game, where one agent's gain is another agent's loss. A GAN learns to generate new data with the same statistics as the training set. See [143] for further details. 31, 34

generative pre-trained transformer (GPT) A family of machine learning models based on the transformer architecture [383] that are pre-trained through SELF-SUPERVISED LEARNING on large data sets of unlabelled text. This is the current predominant architecture for large language models. 34

graph neural network A neural network designed to process graph-structured data. GNNs perform optimizable transformations on graph attributes (e.g., nodes, edges, global context) while preserving graph symmetries such as permutation invariance. GNNs utilize a "graph-in, graph-out" architecture that takes an input graph with information and progressively transforms it into an output graph with the same connectivity as that of the input graph. 31

H

hidden Markov model A Markov model in which the system being modeled is assumed to be a Markov process with unobservable states. The model provides an observable process whose outcomes are influenced by the outcomes of a Markov model in a known way. An HMM can be used to describe the evolution of observable

联邦学习一种机器学习类型，其中模型以去中心化方式使用多个数据源进行训练，而不将数据汇集或集中到任何中心位置。联邦学习允许实体或设备通过交换模型更新来协作训练全局模型，而无需直接共享各实体所控制的数据。5, 31

前馈神经网络 一种人工神经网络，其中各层之间的连接节点从一层到下一层，且不形成循环。31

微调将预训练模型适配以执行特定任务或专门化于特定领域的过程。此阶段紧随初始预训练阶段，涉及在任务特定数据上进一步训练模型。这通常是一项监督学习任务。37

微调规避通过微调移除模型拒绝行为或其他模型级别安全干预。41

形式化方法一种数学上严谨的技术，用于规范、开发和以及软件系统的验证。18

基础模型在生成式人工智能中，使用自监督学习在广泛数据上训练的模型，可通过微调等方式适配各种下游任务[311]。111

功能性攻击一种针对某一领域中一组数据优化的对抗攻击，而非每个数据点。13、23

G

生成对抗网络一种机器学习框架，其中两个神经网络络以零和博弈的形式相互竞争，一方所得即另一方所失。GAN 学习生成与训练集具有相同统计特性的新数据。详见[143]。31, 34

生成式预训练变换器(GPT)一个机器学习模型系列，基于变换器架构[383]的机器学习模型，通过自监督学习在大量无标注文本数据集上进行预训练。这是当前大语言模型的主流架构。34

图神经网络一种设计用于处理图结构数据的神经网络。GNN对图属性(例如节点、边、全局上下文)执行可优化的变换，同时保持图对称性(如置换不变性)。GNN 采用"图入、图出"架构，接收带有信息的输入图，并逐步将其变换为与输入图具有相同连通性的输出图。31

H

隐马尔可夫模型一种马尔可夫模型，其中被建模的系统被假定为具有不可观测状态的马尔可夫过程。该模型提供一个可观测过程，其结果以已知方式受马尔可夫模型结果的影响。HMM 可用于描述可观测

events that depend on internal factors that are not directly observable. In machine learning, it is assumed that the internal state of a model is hidden but not its hyperparameters. 31

I

indirect prompt injection A type of PROMPT INJECTION executed through RESOURCE CONTROL rather than through user-provided input as in a DIRECT PROMPT INJECTION. 39–41, 50

integrity violation In the AML context, an AI system being forced to misperform against its intended objectives, producing outputs or predictions that align with the attacker's objective. 6, 40

J

jailbreak A DIRECT PROMPTING ATTACK intended to circumvent restrictions placed on model outputs, such as circumventing refusal behaviour to enable misuse. 34, 38, 42, 43, 52

L

label flipping A type of data poisoning attack in which an adversary is restricted to changing the training labels. 20

label limit A capability with which an attacker does not control the labels of training samples in supervised learning. 8

logistic regression A type of linear classifier that predicts the probability of an observation being part of a class. 5

M

machine unlearning A technique that involves selectively removing the influences of specific training data points from a trained machine learning model, such as to remove unwanted capabilities or knowledge in a foundation model, or to enable a user to request the removal of their records from a model. Efficient approximate unlearning techniques may not require retraining the ML model from scratch. 33

membership-inference attack A data privacy attack to determine whether a data sample was part of the training set of a machine learning model. 7, 28

misuse enablement In the AML context, a circumvention of technical restrictions imposed by the AI system's owner on its use, such as restrictions designed to prevent a GenAI system from producing outputs that could cause harm to others. 40

model control A capability with which an attacker can control the machine learning model parameters. 8, 37, 41, 111

model extraction A type of privacy attack that extracts details of the model architecture and/or parameters. 7, 28, 31, 40, 41, 47

依赖于无法直接观测的内部因素的事件。在机器学习中，假定模型的内部状态是隐藏的，但其超参数不是。31

I

间接提示注入一种通过资源控制执行的提示注入
而不是像直接提示注入那样通过用户提供的输入。39-41, 50

完整性破坏在对抗性机器学习语境中，指人工智能系统被迫违背其预期目标运行，
产生与攻击者目标一致的输出或预测。6, 40

J

越狱一种直接提示攻击，旨在规避对模型
输出施加的限制，例如规避拒绝行为以促成滥用。34, 38, 42, 43, 52

L

标签翻转一种数据投毒攻击，其中攻击者被限制为只能更改
训练标签。20

标签限制一种能力，攻击者无法控制训练样本的标签
ples在监督学习中的应用。8

逻辑回归一种线性分类器，用于预测观测值的概率
属于某个类别的一部分。5

M

机器遗忘一种技术，涉及从已训练的机器学习模型中选择性地移除特定训练数据点的影响，
例如移除基础模型中不需要的能力或知识，或使用户能够请求从其记录中移除其数据。高效的近似遗忘技术可能不需要从头重新训练机器学习模型。33

成员推断攻击一种数据隐私攻击，用于确定某个数据样本
曾是机器学习模型训练集的一部分。7, 28

滥用促成在对抗性机器学习语境中，指规避人工智能系统所有者对其使用施加的技术限制，
例如旨在防止生成式人工智能系统产生可能对他人造成伤害的输出的限制。40

模型控制一种能力，攻击者可以控制机器学习模型
参数。8, 37, 41, 111

模型萃取一种隐私攻击，提取模型架构的细节
和/或参数。7, 28, 31, 40, 41, 47

model poisoning A POISONING ATTACKS which operates through MODEL CONTROL. 5, 6, 37, 41, 111

model privacy attacks An attack against machine learning models to extract sensitive information about the model. 7

multimodal models A model that processes and relates information from multiple sensory modalities that each represent primary human channels of communication and sensation, such as vision and touch. 58

O

out-of-distribution Data that was collected at a different time and possibly under different conditions or in a different environment than the data collected to train the model. 56

P

poisoning attacks Adversarial attacks in which an adversary interferes with a model during its TRAINING STAGE, such as by inserting malicious training data (DATA POISONING) or modifying the training process itself (MODEL POISONING). 5, 108, 111

pre-training A component of the TRAINING STAGE in which a model learns general patterns, features, and relationships from vast amounts of unlabeled data, such as through SELF-SUPERVISED LEARNING. Pre-training can equip models with knowledge of general features or patterns which may be useful in downstream tasks (see FOUNDATION MODEL), and can be followed with additional training or fine-tuning that specializes the model for a specific downstream task. 37

privacy compromise In the AML context, the unauthorized access of restricted or proprietary information that is part of an AI system, including information about a model's training data, weights or architecture; or sensitive information that the model accesses such as the knowledge base of a GenAI RETRIEVAL-AUGMENTED GENERATION (RAG) application. 7, 40

prompt extraction An attack that tries to divulge the system prompt or other information in the context of a large language model that would normally be hidden from a user. 38, 41

prompt injection An attack which exploits the concatenation of untrusted input with a prompt constructed by a higher-trust party such as the application designer. 38, 41, 108, 110

property inference A data privacy attack that infers a global property about the training data of a machine learning model. 7

Q

query access A capability with which an attacker can issue queries to a trained machine learning model and obtain predictions or generations. 8, 40, 108

R

模型投毒一种通过模型控制实施的投毒攻击。5, 6, 37, 41, 111

模型隐私攻击一种针对机器学习模型、提取敏感信息的攻击关于模型的信息。7

多模态模型一种处理并关联来自多种感官模态信息的模型，每种模态分别代表人类主要的交流与感知通道，例如视觉和触觉。58

O

分布外在与训练模型所用数据不同的时间、可能不同的条件，或在不同于训练模型所收集数据的环境中。56

P

投毒攻击一种对抗攻击，攻击者在模型的训练阶段进行干扰，例如插入恶意训练数据(数据投毒)或修改训练过程本身(模型投毒)。5, 108, 111

预训练训练阶段的一个组成部分，模型从大量无标签数据中学习通用模式、特征和关系，例如通过自监督学习。预训练可以使模型具备通用特征或模式的知识，这些知识可能在下游任务中有用(参见基础模型)，并且可以随后进行额外的训练或微调，使模型专门化以用于特定的下游任务。37

隐私泄露在对抗性机器学习语境中，指未经授权访问属于人工智能系统的受限或专有信息，包括关于模型训练数据、权重或架构的信息；或模型访问的敏感信息，例如生成式人工智能检索增强生成(RAG)应用的知识库。7, 40

提示萃取一种试图泄露系统提示或其他信息的攻击在大语言模型的语境中，这些信息通常会对用户隐藏。38, 41

提示注入一种利用不可信输入与由更高信任方(如应用设计者)构造的提示相拼接的攻击。38, 41, 108, 110

属性推断一种数据隐私攻击，用于推断关于训练数据的全局属性机器学习模型的数据。7

Q

查询访问一种能力，攻击者可借此向训练好的机器学习模型发起查询学习模型并获取预测或生成结果。8, 40, 108

R

receiver operating characteristic (ROC) A curve that plots the true positive rate versus the false positive rate for a classifier. 107

red teaming in the AI context, means a structured testing effort, often adopting adversarial methods, to find flaws and vulnerabilities in an AI system, including unforeseen or undesirable system behaviors or potential risks associated with the misuse of the system. [366]. 60

regression A type of supervised machine learning model that is trained on data, including numerical labels (i.e., response variables). Types of regression algorithms include linear regression, polynomial regression, and various non-linear regression methods. 5

reinforcement learning A type of machine learning in which a model learns to optimize its behavior according to a reward function by interacting with and receiving feedback from an environment. 5

resource control A capability in which an attacker controls one or more external resources consumed by a machine learning model at inference time, particularly for GenAI systems such as retrieval-augmented generation applications. 41, 50, 108, 110

retrieval-augmented generation (RAG) A type of GenAI system in which a model is paired with a separate information retrieval system (or ”knowledge base”). Based on a user query, the RAG system identifies relevant information within the knowledge base and provides it to the GenAI model in context for the model to use in formulating its response. RAG systems allow the internal knowledge of a GenAI model to be modified without the need for retraining. 1, 35, 37, 38, 40, 46, 50, 111

rowhammer attack A software-based fault-injection attack that exploits dynamic random-access memory disturbance errors via user-space applications and allows the attacker to infer information about certain victim secrets stored in memory cells. Mounting this attack requires the attacker to control a user-space unprivileged process that runs on the same machine as the victim’s machine learning model. 31

S

self-supervised learning A type of machine learning that relies on generating implicit labels from unstructured data rather than relying on explicit, human-created labels. Self-supervised learning tasks are constructed to allow the true labels to be automatically inferred from the training data (enabling the use of large-scale training data) and to require models to capture essential features or relationships within the data to solve them. For example, a common self-supervised learning task is providing a model with partial data with the task to accurately generate the remainder. 109, 111

semi-supervised learning A type of machine learning in which a small number of training samples are labeled, while the majority are unlabeled. 5

shadow model A model that imitates the behavior of the target model. The training datasets and the truth about membership in these datasets are known for these models.

受试者工作特征曲线(ROC)一条绘制真正例率与
分类器的假阳性率。107

红队测试在人工智能语境中，指一种结构化的测试工作，通常采用对抗性
方法，以发现人工智能系统中的缺陷和漏洞，包括未预见的或不良的系统行为，或与系统滥用相关的潜在风险。[366]。60

回归 一种监督机器学习模型，它在包含
数值标签(即响应变量)的数据上进行训练。回归算法的类型包括线性回归、多项式回归以及各种非线性回归方法。5

强化学习一种机器学习类型，模型通过与环境的交互并接收反馈，
通过与环境的交互并接收反馈，按照奖励函数行事。5

资源控制一种能力，攻击者控制机器学习模型在推理时消耗的一个或多个外部资源，
尤其针对检索增强生成应用等 GenAI 系统。41, 50, 108, 110

检索增强生成(RAG)一种 GenAI 系统类型，其中模型与
一个独立的信息检索系统(或“知识库”)配对。基于用户查询，RAG 系统在知识库中识别相关信息，并将其作为上下文提供给 GenAI 模型，供模型在组织其响应时使用。RAG 系统允许在无需重新训练的情况下修改 GenAI 模型的内部知识。1, 35, 37, 38, 40, 46, 50, 111

Rowhammer 攻击一种基于软件的故障注入攻击，通过用户空间应用程序利用动态随机
存取存储器扰动错误，使攻击者能够推断存储在存储单元中的某些受害者秘密信息。实施此攻击需要攻击者控制一个运行在与受害者机器学习模型同一台机器上的用户空间非特权进程。

31

S

自监督学习一种机器学习类型，依赖于生成隐式标
从非结构化数据中学习，而不是依赖显式的、人工创建的标签。自监督学习任务被构建为允许从训练数据中自动推断真实标签(从而能够使用大规模训练数据)，并要求模型捕获数据中的基本特征或关系以解决这些任务。例如，一个常见的自监督学习任务是向模型提供部分数据，任务是准确生成剩余部分。109, 111

半监督学习一种机器学习类型，其中少量训练
样本已标注，而大多数样本未标注。5

影子模型一种模仿目标模型行为的模型。训练数据集
而这些模型中，关于成员资格在这些数据集中的真相是已知的。

Typically, the attack model is trained on the labeled inputs and outputs of the shadow model. 25

side channel Allows an attacker to infer information about a secret by observing the non-functional characteristics of a program (e.g., execution time or memory) or measuring or exploiting the indirect coincidental effects of the system or its hardware (e.g., power consumption variation, electromagnetic emanations) while the program is executing. Most commonly, such attacks aim to exfiltrate sensitive information, including cryptographic keys. 31

source code control A capability with which an attacker controls the source code of a machine learning algorithm. 8

supervised learning A type of machine learning in which a model learns to predict explicit (often human-generated) labels or output values for data. 5

support vector machines Models that implement a decision function in the form of a hyperplane that serves to separate (i.e., classify) observations that belong to one class from another based on patterns of information about those observations (i.e., features). 5, 6, 31

system prompt Application-specific instructions provided in-context to a GenAI system by the model developer or application designer. System prompts are typically prepended to other input, and may be higher-trust than other forms of input. 38, 43, 46, 52

T

targeted poisoning attack A poisoning attack that changes the prediction on a small number of targeted samples. 6, 42

testing data control A capability with which an attacker controls the testing data input to the machine learning model. 8

training data control A capability in which an attacker controls some or all of the training data of a machine learning model. 7, 40

training data extraction The ability of an attacker to extract the training data of a generative model by prompting the model with specific inputs. 7, 46

training stage The stage of a machine learning pipeline in which a model learns parameters that minimize its error against an objective function based on training data. 5, 37, 111

trojan In the machine learning context, a malicious modification to a model that is difficult to detect, may appear harmless, but that can alter the intended function of the system upon a signal from an attacker to cause a malicious behavior desired by the attacker. For Trojan attacks to be effective, the trigger must be rare in the normal operating environment so that it does not affect the normal effectiveness of the AI and raise the suspicions of users. In the machine learning context, trojan may be used interchangeably with backdoor pattern. 2

U

通常，攻击模型在影子模型的带标签输入和输出上进行训练。25

侧信道 允许攻击者通过观察非
功能性特征(如执行时间或内存)或在程序执行期间测量或利用系统或其硬件的间接巧合效应(如功耗变化、电磁辐射)来推断关于秘密的信息。最常见的是，此类攻击旨在窃取敏感信息，包括密码密钥。31

源代码控制一种能力，攻击者可借此控制机器学习
机器学习算法。8

监督学习一种机器学习类型，其中模型学习预测显式
(通常是人工生成的)数据标签或输出值。5

支持向量机以超平面形式实现决策函数的模型，
该超平面用于基于关于观测的信息模式(即特征)将属于一个类别的观测与另一个类别分离(即分类)。5, 6, 31

系统提示由模型开发者或应用设计者以上下文方式提供给 GenAI 系统的
应用特定指令。系统提示通常被前置到其他输入之前，并且可能比其他形式的输入具有更高的信任级别。38, 43, 46, 52

T

定向投毒攻击一种投毒攻击，其改变对少量
目标样本数量。6, 42

测试数据控制一种能力，攻击者借此控制输入到
机器学习模型。8

训练数据控制一种能力，攻击者借此控制部分或全部训练
机器学习模型的数据。7, 40

训练数据提取攻击者提取生成式模型训练数据的能力
通过向模型提供特定输入来提示模型，从而生成生成式模型的输出。7, 46

训练阶段机器学习流水线中模型学习参数的阶段，
这些参数使其基于训练数据的目标函数误差最小化。5, 37, 111

木马在机器学习语境中，指对模型的一种恶意修改，难以
被检测，可能看似无害，但可在收到攻击者信号时改变系统的预期功能，以产生攻击者期望的恶意行为。为使木马攻击有效，触发器必须在正常运行环境中罕见，从而不影响人工智能的正常效能并引起用户怀疑。在机器学习语境中，木马可与后门模式互换使用。2

U

unsupervised learning A type of machine learning in which a model learns based on patterns in unlabeled data, such as learning a function to cluster or group data points. 5

无监督学习一种机器学习类型，模型基于无标签数据中的模式进行学习，例如学习对数据点进行聚类或分组的函数。 5

How this file was made

This is a complete Chinese translation of NIST AI 100-2e2025, Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. Tables, captions, headers, footers and numbering all keep their original positions. Nothing has been cut.

If you have documents of your own

- Contracts, tenders, standards, patents, reports: original layout kept, CNY 0.15/page
- PDF to Word: free up to 100 pages, CNY 0.10/page beyond that
- Scanned or image-only PDFs: OCR to Word, free
- No sign-up, no login: upload and download

Open the address below and try it with your own file:

https://www.ivyivy.com.cn/nist-ai-100-2-zh

On-premise deployment available (subject to model and compliance review)

Machine translation, provided for reading reference only. For professional judgement and compliance citation, please refer to the original English source.

This page is a translator's note; it is not part of the original document.

这份文件是怎么做出来的

本译本是《对抗性机器学习：攻击、缓解措施与分类（AI 100-2）》的中文全译本，共 127 页。表格、图注、页眉页脚、编号层级都按原文位置一一还原，未做任何删节。

你手上如果也有英文文件要处理

- 合同、标书、标准、专利、报告：翻译后保留原版式，0.15 元/页
- PDF 转 Word：100 页内免费，超出 0.1 元/页
- 扫描件、图片版 PDF：先 OCR 再转 Word，免费
- 不用注册、不用登录，传上去直接下载

打开下面这个地址，把你自己的文件传上去试试：

https://www.ivyivy.com.cn/nist-ai-100-2-zh

支持私有化部署（需按贵方模型与合规要求评估）

本译本由机器翻译生成，仅供阅读参考；专业技术判断与合规引用，请以英文原文为准。

本页为译制说明，不属于原文件内容。