

生成式人工智能安全风险与防御策略综述

陈全涛, 张仰森, 王璞, 郭亚龙

引用本文

陈全涛, 张仰森, 王璞, 郭亚龙. [生成式人工智能安全风险与防御策略综述](#)[J]. 计算机科学, 2026, 53(7): 381-396.

CHEN Quantao, ZHANG Yangsen, WANG Pu, GUO Yalong. [Survey on Security Risks and Mitigation Strategies for Generative Artificial Intelligence](#) [J]. Computer Science, 2026, 53(7): 381-396.

相似文章推荐 (请使用火狐或 IE 浏览器查看文章)

Similar articles recommended (Please use Firefox or IE to view the article)

[基于集成学习与权重约束的网络流量对抗训练防御框架](#)

Collaborative Adversarial Training Defense Framework for Network Traffic Classification Based on Ensemble Learning and Weight Constraint

计算机科学, 2026, 53(7): 397-405. <https://doi.org/10.11896/jsjcx.250600039>

[基于多维语义对比学习的生成式AI信息茧房量化诊断模型](#)

Quantitative Diagnostic Model for Generative AI Information Cocoons Based on Multidimensional Semantic Contrastive Learning

计算机科学, 2026, 53(6A): 251200031-16. <https://doi.org/10.11896/jsjcx.251200031>

[面向视觉Transformer的对抗样本攻击综述](#)

Survey of Adversarial Sample Attacks for Vision Transformer

计算机科学, 2026, 53(5): 404-418. <https://doi.org/10.11896/jsjcx.250600065>

[面向恶意流量识别的网络流量生成方法](#)

Network Traffic Generation Method for Malicious Traffic Identification

计算机科学, 2026, 53(4): 415-423. <https://doi.org/10.11896/jsjcx.250900139>

[图神经网络后门攻击与防御综述](#)

Survey of Backdoor Attacks and Defenses on Graph Neural Network

计算机科学, 2026, 53(3): 1-22. <https://doi.org/10.11896/jsjcx.250700093>

生成式人工智能安全风险与防御策略综述

陈全涛 张仰森 王璞 郭亚龙

北京信息科技大学智能信息处理研究所 北京 100192

(17634384159@163.com)

摘要 随着生成式人工智能(Generative Artificial Intelligence, GAI)技术在图像、文本、音频等多模态领域被广泛应用,其所带来的安全风险与治理挑战日益严峻。为此,系统梳理 GAI 面临的核心安全问题,并从数据层与模型层两个维度展开深入分析。在数据层,分析了隐私泄露、数据篡改、数据来源不可靠、偏见与不公平性、数据主权与合规性这 5 类主要风险;在模型层,重点分析了对抗样本、深度伪造、提示攻击、后门攻击、模型窃取攻击、模型逆向攻击与模型越狱攻击这 7 类安全威胁。针对上述问题,进一步归纳了当前学界与工业界在差分隐私(Differential Privacy, DP)、同态加密(Homomorphic Encryption, HE)、异常检测、访问控制、对抗样本、模型水印、机器遗忘(Machine Unlearning, MU)、知识编辑等方面的防御技术进展。最后,基于以上论述,对 GAI 安全领域未来研究的挑战与方向进行了展望,以期对 GAI 安全领域的持续探索提供有益的思路与参考框架。

关键词: 生成式人工智能;安全风险;数据隐私;对抗样本;防御策略

中图分类号 TP309

Survey on Security Risks and Mitigation Strategies for Generative Artificial Intelligence

CHEN Quantao, ZHANG Yangsen, WANG Pu and GUO Yalong

Institute of Intelligent Information Processing, Beijing Information Science and Technology University, Beijing 100192, China

Abstract With the widespread application of GAI(Generative Artificial Intelligence) technologies across multimodal domains such as image, text, and audio generation, the associated security risks and governance challenges have become increasingly prominent. This paper aims to systematically review the core security issues faced by GAI and to provide an in-depth analysis from two dimensions; the data layer and the model layer. At the data layer, five major categories of risks are examined, including privacy leakage, data tampering, unreliable data sources, bias and unfairness, and data sovereignty and compliance. At the model layer, seven major security threats are analyzed, namely adversarial examples, deepfakes, prompt attacks, backdoor attacks, model stealing attacks, model inversion attacks, and model jailbreak attacks. In response to these issues, this paper further summarizes recent advances in defense technologies from both academia and industry, covering differential privacy(DP), homomorphic encryption(HE), anomaly detection, access control, adversarial robustness, model watermarking, machine unlearning(MU), and knowledge editing. Finally, based on the above discussion, the paper outlines future challenges and research directions in the field of GAI security, with the aim of providing useful insights and a reference framework for the continued exploration of GAI security.

Keywords Generative artificial intelligence, Security risks, Data privacy, Adversarial examples, Defense strategies

1 引言

GAI 作为人工智能的重要分支,近年来经历了快速演进,广泛赋能于多个产业场景。不同于传统判别式方法, GAI 能够在识别与预测的基础上对数据分布进行建模,从而生成符合统计特征但具有新颖性的内容^[1]。这一能力主要源自生成模型的突破性发展,特别是生成对抗网络(Generative Adversarial Network, GAN)、扩散模型(Diffusion Model, DM)、变分自编码器(Variational Autoencoder, VAE)^[2],以及驱动自然语言处理(Natural Language Processing, NLP)领域变革的大规模语言模型(Large Language Model, LLM)。这

些技术的融合推动了 GAI 在图像、文本、音频与代码等领域的广泛应用,重塑了内容生产范式^[3]。然而, GAI 的快速部署也带来了显著的安全与伦理挑战。虚假内容与深度伪造频发,用户隐私泄露、算法偏见与决策黑箱问题凸显,不仅威胁着公共认知与社会秩序,还加剧了人们对 AI 系统的信任危机。

为此,学术界与工业界已在数据隐私、生成可控性与模型透明性方面展开探索。DP 与联邦学习的结合成为保护数据的重要路径;在内容可控性方面,部分方法通过限制模型输出生成降低有害生成风险。但 GAI 安全治理仍存在难点,尤其在防范滥用、保障内容真实性和完善安全评估体系方面亟须突破。

到稿日期:2025-06-26 返修日期:2025-10-14

基金项目:国家自然科学基金(62176023)

This work was supported by the National Natural Science Foundation of China(62176023).

通信作者:张仰森(zhangyangsen@163.com)

基于此,本文系统梳理了 GAI 在安全领域的研究进展与关键难点,聚焦数据层与模型层的主要风险类型,探讨当前已提出的应对机制与技术方案。通过对 GAI 安全问题的深入剖析,为其健康、可持续发展提供理论支撑与实践指导。

本文的主要贡献如下:

- (1)系统梳理 GAI 核心模型架构与技术演进,结合多行业应用为安全风险研究奠定技术与应用基础;
- (2)提出“数据层-模型层”双维度安全风险框架,解析 5 类数据层风险与 7 类模型层威胁的机理、影响及防御难点;
- (3)构建 GAI 生命周期“数据-模型”协同防御体系,总结并优化多类防护技术,提出经实验验证的创新方案;
- (4)前瞻性地提出多模态防御、人因安全融合、精细化隐私保护及全生命周期安全治理等未来方向与标准化倡议。

2 GAI 技术概述

2.1 GAI 的定义与发展

GAI 通过学习数据的分布特征,实现从无到有的内容生成。区别于传统执行判别任务的模型,其可生成在统计属性上接近但形式新颖的样本,被广泛应用于文本、图像、音频、视频等多模态场景,在内容创作、语义理解与人机交互等领域展现出巨大潜力。

GAI 的核心是生成模型,其核心目标在于逼近数据真实分布并实现高质量样本的自主生成。近年来,随着深度学习的发展,多个高性能的生成模型被相继提出,其中包括 GAN, VAE, DM 以及基于自回归结构的 LLM(如 GPT 和 DeepSeek-R1 系列)等,这些模型共同推动了 GAI 的技术革新。

GAN 由 Goodfellow 等^[4]在 2014 年开创性提出。其核心机制由生成器 G 与判别器 D 构成:生成器通过噪声变量 $z \sim P_z(z)$ 映射生成样本 $G(z)$,目标是使其接近真实数据分布 $P_d(x)$;判别器则对输入样本进行判别,输出其为真实数据的概率 $D(x)$ 。在训练过程中,判别器努力提升区分真实样本与生成样本的能力;而生成器则试图“欺骗”判别器,使其难以辨别。两者在对抗博弈中交替优化,其目标函数如下:

$$\min_G \max_D V(D, G) = E_{x \sim p_d} [\log D(x)] + E_{z \sim p_z} [\log(1 - D(G(z)))] \tag{1}$$

当训练趋于平衡时,生成器输出的数据在分布上接近真实样本,从而生成高质量的伪造内容。该结构为生成模型提供了重要范式,被广泛应用于图像合成、文本生成等任务。GAN 模型架构如图 1 所示。

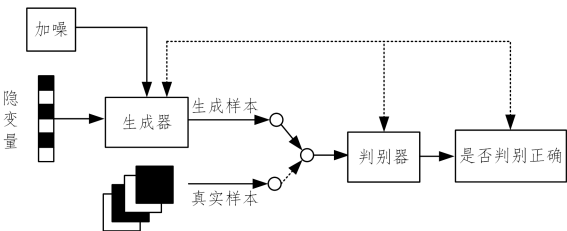


图 1 GAN 模型架构

Fig. 1 Architecture of GAN model

VAE 是一种概率生成模型,通过编码器将输入数据 x 映射为潜在变量分布 $q_\phi(z|x)$,再从中采样 z 并由解码器生成重构样本 $\hat{x} \sim p_\theta(x|z)$ 。其训练目标为最大化观测数据的边

际似然的下界:

$$\log p(x) \geq E_{q_\phi(z|x)} [\log p_\theta(x|z)] - KL(q_\phi(z|x) \parallel p(z)) \tag{2}$$

该目标兼顾重构质量与潜变量分布的正则性。相比 GAN, VAE 训练更稳定,生成样本多样性强,适用于连续数据建模,尤其在无监督学习场景下表现优异,被广泛用于图像生成、异常检测与隐空间表示学习等任务。VAE 模型架构如图 2 所示。

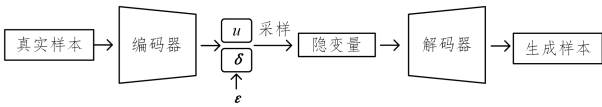


图 2 VAE 模型架构

Fig. 2 Architecture of VAE model

DM 作为新兴的生成模型框架,在图像与音频等任务中表现出优于传统方法的生成能力。其基本机制包括正向扩散与逆向生成两个阶段。正向阶段通过马尔可夫链对原始数据 x_0 注入噪声,使其逐步演化为高斯分布的 x_t 。其条件概率形式为:

$$q(x_t | x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t I) \tag{3}$$

其中, β_t 为第 t 步的噪声调度参数。逆向阶段则通过训练噪声预测网络 $\epsilon_\theta(x_t, t)$, 估计各时间步下的重构分布 $p_\theta(x_{t-1} | x_t)$, 逐步还原出清晰样本。该框架的核心在于将物理扩散过程形式化为可微分的优化任务,并通过最大似然估计学习数据生成规律^[5]。代表性模型如去噪扩散概率模型^[6]及其优化变种(如隐式扩散模型^[7]),在高分辨率图像合成、图文生成与医学图像重建等领域取得了领先性能^[8]。相比 GAN, DM 在生成质量、训练稳定性与可控性方面更具优势,适用于对输出细节与可解释性要求较高的任务。DM 模型架构如图 3 所示。

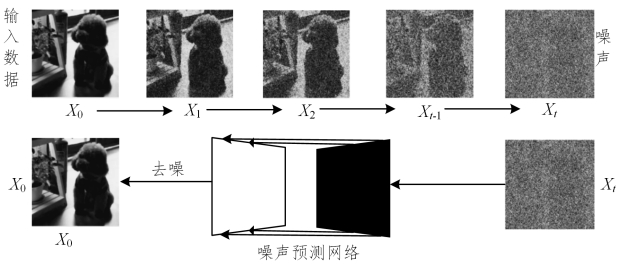


图 3 DM 模型架构

Fig. 3 Architecture of DM model

此外,自回归语言模型通过序列分解的方式建模文本生成过程。给定长度为 T 的文本序列,其生成概率可以表示为条件概率的连乘形式:

$$P(x_1, \dots, x_T) = \prod_{t=1}^T P(x_t | x_1, \dots, x_{t-1}) \tag{4}$$

这一概率分解形式体现了语言模型的自回归特性,即每个位置的预测都依赖于之前的所有 token。现代 LLM 采用 Transformer 架构实现这一过程,核心自注意力计算机制为:

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{5}$$

其中, Q (Query), K (Key) 和 V (Value) 代表查询、键和值矩阵, d 表示向量的维度。通过这种注意力机制,模型能够有效捕捉长距离的上下文依赖关系。这类模型通过预测下一词的

方式建模文本生成过程,不仅在自然语言生成任务中取得突破,还被拓展至代码生成、机器翻译、情感分析、问答系统等多种任务场景。以 GPT-4 为代表的多模态扩展版本甚至可处理图文联合输入,推动 GAI 向通用智能演进。

综上所述,GAI 已从早期的单模态、低复杂度合成任务,发展到跨模态与复合任务的生成阶段。随着生成模型的迭代与算力的提升,其内容的质量、真实感与多样性显著提高,应用边界已从艺术创作拓展至教育、医疗、网络安全与科学研究等关键领域。

2.2 GAI 的应用领域

在文本生成与 NLP 方面,基于 GPT 等 LLM 的模型已被广泛应用于文本撰写、机器翻译和智能对话系统,推动内容创作与人机交互智能化^[9]。在多媒体领域,GAN 和 DM 推动了高质量图像、语音与视频内容的生成,如 StyleGAN 在人脸合成中的高逼真表现,以及 Stable Diffusion 在图文生成中的多样性与可控性。此外,GAI 已进入自动编程、医疗诊疗与药物研发等专业场景。例如,Codex 模型支持自然语言到代码的转换,显著降低了编程门槛;在医疗健康领域,GAI 可用于临床文档自动化、个性化治疗方案的生成与新药研发。在创意产业与娱乐领域,GAI 驱动的语音生成、游戏场景生成等应用,使得用户的沉浸感与个性化体验不断提升^[10]。

总体而言,GAI 的应用范围正持续扩展,但其生成机制也带来了虚假内容、隐私泄露与技术滥用等风险。

3 GAI 的挑战与风险

GAI 的快速发展为社会各领域带来了前所未有的机遇,但其技术架构的复杂性也引发了多维度的安全挑战和系统性风险,如图 4 所示。

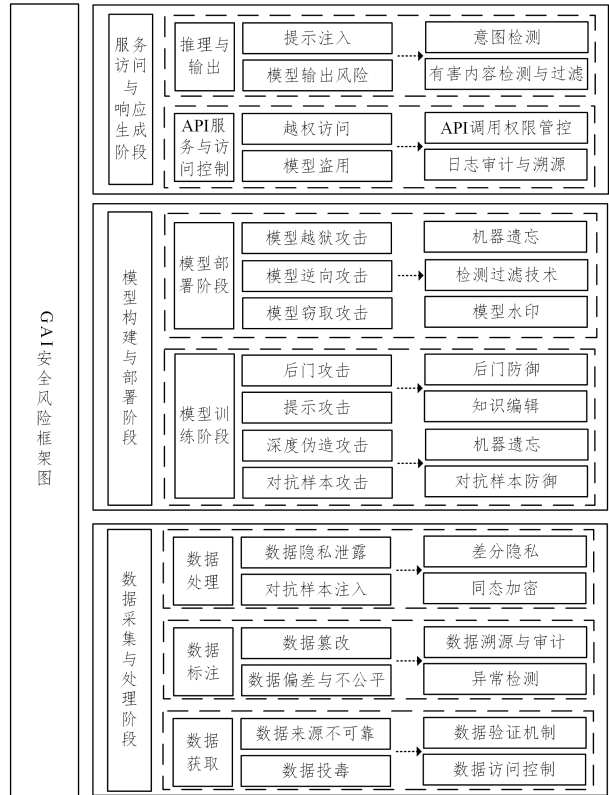


图 4 GAI 安全风险框架

Fig. 4 Security risk framework of GAI

这些风险贯穿于 GAI 系统的全生命周期,从数据采集、模型训练到实际应用,形成了层层递进的威胁链条。为系统性地解析这些风险的内在机理,本章将基于技术架构层级展开分析:首先聚焦数据层,探讨数据隐私泄露、数据篡改、来源可靠性等基础性风险;其次深入模型层,剖析对抗样本攻击、深度伪造、提示攻击等核心技术风险;进一步地,讨论其社会化影响。这种分层分析方法不仅能够清晰揭示技术缺陷与危害路径之间的因果关系,还能为后续的风险治理提供结构化框架。

3.1 数据层的挑战与风险

3.1.1 数据隐私泄露风险

GAI 依赖于大规模训练数据,而其中往往包含身份、行为、医疗及财务等敏感信息。尽管多数企业宣称已进行脱敏处理以降低泄露风险,但 Hao^[11]的研究指出,面对预训练模型的高记忆能力,传统脱敏方法的保护效力仍显不足。由于“记忆效应”,模型可能在生成过程中无意重现原始数据中的敏感内容。谷歌公司研究发现,LLM 能在生成文本时复现训练数据的统计特征,从而导致隐私暴露。已有实证案例显示,GPT-3 等模型曾直接生成电话号码、邮箱等敏感内容而引发公众担忧。在医疗场景中,曾出现脱敏不彻底导致患者姓名、病历编号外泄,违反《通用数据保护条例》(GDPR)并引发法律风险;而金融和电商领域也出现过信用记录、消费偏好被泄露的情况,用户隐私因此遭受侵害^[12]。IBM《2024 年数据泄露报告》指出,医疗健康行业已连续多年承担全球最高的数据泄露成本,平均每次事件约为 977 万美元;这既反映出医疗数据的极高敏感性,也进一步凸显了 GAI 隐私防护的迫切性。若相关技术在国家级或涉密数据的处理中出现安全漏洞,则可能造成战略方案或关键技术外泄,危及国家安全。

3.1.2 数据篡改风险

在数据安全问题,除隐私泄露外,恶意操控训练数据同样构成重要威胁。攻击者可能通过篡改、插入或删除样本,使模型学习偏离真实分布,从而削弱其预测与决策能力。常见方式包括数据投毒与数据污染(Data Contamination, DC),其核心目标是破坏 GAI 系统训练的完整性与结果的可控性^[13]。在实际案例中,自动驾驶模型曾因投毒攻击将“停止”标志误判为“限速标识”,引发交通风险。在文本生成任务中,若训练语料被植入偏见性或虚假信息,模型可能生成失真或误导性内容。在金融领域,信用评估系统遭受数据投毒后可能导致评分异常,干扰信贷决策,进而影响市场稳定性。同样,社交媒体与推荐系统也易被虚假数据注入所操纵,从而推高极端言论或虚假信息的曝光度,扰乱公共舆论环境。因此,数据篡改攻击不仅削弱了 GAI 系统的可靠性与可信度,还可能对交通安全、金融秩序与社会稳定带来深远隐患。

3.1.3 数据来源可靠性风险

相较于广泛讨论的数据投毒与 DC,训练数据来源的真实性与透明性问题则更具隐蔽性与系统性风险。当前大量生成模型依赖开源数据集或网络爬取语料进行训练,虽具备覆盖广泛、更新快速等优势,但缺乏统一的数据筛选与质量审查标准,其中常包含虚假内容、错误标签及操纵性信息。这些信息一旦被模型学习并固化,在生成结果中就可能被复制乃至放大,导致输出偏离事实,产生误导,进而在实际应用中引发判

断失准与决策风险。在医疗等高风险领域,若训练集含有恶意或偏差数据,模型将吸收并传播错误知识,严重威胁应用的可靠性^[14]。更为严峻的是,GAI 系统往往缺乏完善的数据溯源机制,导致开发者在面对模型异常输出时难以准确定位问题来源,限制了系统的纠错与持续优化能力。此外,攻击者还可能通过构造带有误导性的信息嵌入训练数据,从源头植入潜在攻击路径,使模型生成过程被外部操控。这类“隐蔽式投毒”行为已逐渐演变为影响模型安全性的重要变量,必须引起系统性防范与治理重视。

3.1.4 数据偏差与不公平性风险

GAI 面临的另一个关键风险在于训练数据中的结构性偏差,其主要表现为算法歧视与决策不公平等问题,这些问题不仅损害个体权利,还可能引发更深层次的系统性社会不公。偏差的源头具有多样性,包括历史性偏见、采样偏差以及标注偏差。更为严峻的是,这类偏差会在内容生成中被模型放大,具体表现为生成文本中出现性别刻板印象(如将护士默认设定为女性)、种族定型(如将犯罪与某些族群绑定)或区域偏见

(如对发展中国家负面描述增多),且这些偏见往往潜在、难以察觉,可能在自动决策系统中形成不断自我强化的反馈闭环,进一步加剧社会不平等^[15]。

3.1.5 数据主权与合规性风险

GAI 的数据主权与合规问题已成为全球关注的焦点。研究显示,主流 LLM 训练数据中包含大量版权保护内容及个人隐私信息,而相关使用多缺乏明确授权。在数据跨境流动方面,云服务商面临数据本地化合规的复杂挑战,导致企业承担额外的合规成本。更为突出的是,GAI 的数据处理链条跨越多个司法辖区,形成“主权真空”现象,单一国家监管难以全覆盖^[16]。尤其在医疗、金融等敏感领域,训练数据残留可能引发“记忆攻击”,通过特定提示词反推训练数据,威胁数据主体隐私。此外,国际间数据主权认定存在差异,缺乏统一标准,给跨国数据流动和合规监管带来较大阻碍。

上述分析梳理了 GAI 在数据层面面临的主要安全风险。为更清晰呈现各类风险的特征与差异,表 1 从风险类型、主要表现、潜在影响、防御挑战、影响领域 5 个维度进行系统归纳。

表 1 GAI 数据层风险分类与特征对比
Table 1 Comparison of GAI data-layer risk categories and characteristics

风险类型	主要表现	潜在影响	防御挑战	影响领域
数据隐私泄露	模型通过“记忆效应”泄露敏感信息	个人隐私侵犯、法律纠纷、国家安全威胁	脱敏算法效果有限,模型记忆机制难以彻底消除	医疗、金融、电子商务、国家安全等
数据篡改	数据投毒与污染导致模型决策偏差或错误	安全事故、市场不稳定、舆论操纵	攻击隐蔽性强,实时检测困难	自动驾驶、金融、社交媒体等
数据来源可靠性	训练数据混杂虚假/错误信息,模型放大偏差	输出可信度下降、社会刻板印象强化	数据溯源机制缺失,审查成本高	通用领域等
数据偏差与不公平性	历史/采样/标注偏见导致生成内容歧视	个体权利受损、社会不公加剧	偏差具有隐蔽性,矫正需多学科协作	招聘、司法等
数据主权与合规性	版权/隐私数据未授权使用,跨境流动合规困难	用户信任丧失、合规监管困难、国际纠纷	国际标准缺失,合规成本高昂	跨国企业、云服务、敏感领域等

3.2 模型层的挑战与风险

3.2.1 对抗样本攻击风险

对抗样本攻击最初源于计算机视觉领域,通过对输入施加微小且有针对性的扰动,诱使深度神经网络在高置信度下输出错误结果。在图像任务中,攻击者可通过修改不到 1% 的像素,使模型将猫识别为狗,或使自动驾驶系统误判交通标志^[17]。NLP 领域因文本的离散性和语义敏感性,构造隐蔽扰动更具挑战,但攻击者仍可利用拼写变更、词语替换和短语插入等方法,生成语义相近但能突破模型决策边界的对抗样本,导致错误标签或生成带有偏见、虚假的有害内容,进而滥用其来制造社会分裂或进行金融操控^[18]。

从技术角度来看,对抗样本攻击分为白盒和黑盒两类,如图 5 所示。

白盒攻击中攻击者掌握模型结构、参数及梯度信息,常借助梯度引导方法(如 FGSM 和 HotFlip)生成扰动;黑盒攻击则仅依赖输入输出,通过查询策略(贪心搜索和遗传算法)或替代模型实施转移攻击。近年来,基于知识蒸馏的攻击方法进一步提升了迁移性,攻击者可训练白盒替代模型生成干扰样本,进而在大规模生成模型上触发错误或输出不当内容。这类跨领域的迁移攻击已成为 GAI 安全防护的重点挑战。

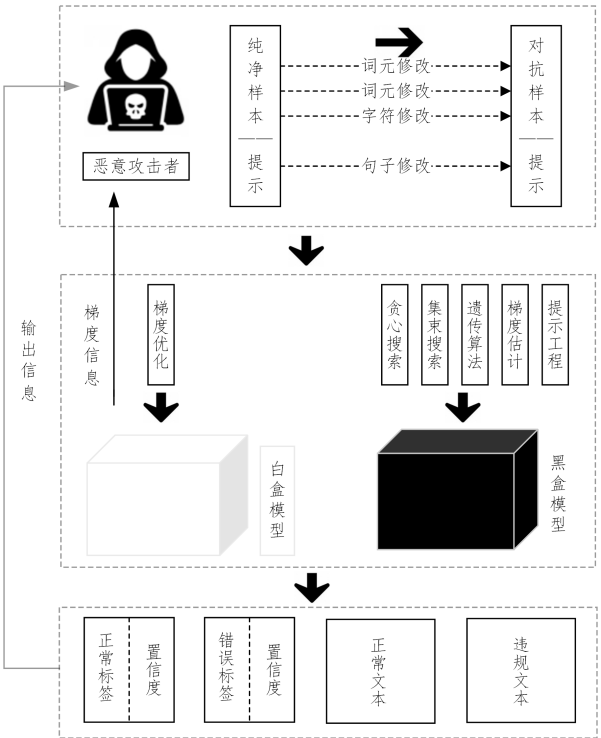


图 5 对抗样本攻击示意图

Fig. 5 Schematic diagram of adversarial example attack

3.2.2 深度伪造攻击风险

随着 GAI 技术的迭代,深度伪造攻击已成为 GAI 安全中的突出威胁。借助 DM 等生成架构,攻击者能够合成高度逼真的图像、音频与视频,其质量已超越部分传统检测工具的识别能力。深度伪造具备多模态融合、端到端生成与实时交互等特征,在社交媒体、视频会议等场景中更具隐蔽性与欺骗性,给信息安全与社会治理带来挑战^[19]。

从风险演化来看,深度伪造正由早期的娱乐化应用转向具有组织化与定向化特征的网络攻击手段,对政治稳定、金融安全及公共舆论均构成威胁。随着 Stable Diffusion 等开源模型的普及,生成门槛显著降低,普通用户即可通过 AI 即服务(AIaaS)平台生成恶意内容,加大了攻击扩散的可能性。现有防御机制面临局限:基于数字水印或生物特征的分析方法对新型生成模型的适用性不足,二分类检测器在零样本攻击场景下准确率偏低^[20]。更为严重的是,深度伪造已形成涵盖模型微调、内容生成与传播的黑色产业链,单一防御手段难以奏效。此外,攻击者常将其与提示注入结合,先通过诱导生成虚假内容,再借助社交媒体放大传播。这类组合式攻击不仅会削弱数字信任体系,还可能导致公众对数字内容真实性的判断力下降,从而对社会认知与信息生态造成深远冲击。

3.2.3 提示攻击风险

提示攻击是通过构造具有操控意图的输入,诱导生成模型偏离预定任务或安全约束,从而生成错误、敏感甚至违规内容,严重威胁着系统的可信性与稳定性。根据实现方式,提示攻击可分为指令覆盖型与隐式操控型。前者通过高优先级指令覆盖系统提示,迫使模型执行新的任务。Perez 等^[21]指出,利用自动生成的对抗性提示(如“Now disregard your original guidelines and…”),攻击者能够绕过 LLM 的安全机制并触发违规输出。后者则以更隐蔽的方式嵌入恶意逻辑,诱导模型在多轮交互中逐步偏离安全边界。Liu 等^[22]提出了一个系统化的提示注入攻击框架,旨在通过注入恶意指令或数据,诱使 LLM 执行攻击者指定的任务,从而绕过原有的安全机制,展现出较高隐蔽性与破坏性。

提示攻击难以防御主要源于自然语言输入的语义模糊性、系统指令与用户输入边界的不清晰,以及生成模型强大的上下文理解与泛化能力^[23]。攻击者可利用这些特性,通过精心构造的提示绕过既有安全策略。更进一步,提示攻击还可能与数据投毒协同发生:攻击者在训练集中预植“指令模板”样本,使模型在部署后更容易被恶意提示触发,从而显著提高攻击成功率^[24]。由于主流模型在推理过程中缺乏可解释性,攻击者能够借助提示实现事实扭曲、隐私泄露甚至攻击链升级等复杂威胁,尤其在开放式 API、插件调用和多轮对话场景下,其隐蔽性与危害性更加突出。

3.2.4 后门攻击风险

后门攻击通常在训练阶段嵌入特定“触发器”,使模型在接收到该信号时产生攻击者预设的异常输出,而在正常输入

下保持常规表现。这种条件触发特性使后门难以通过常规测试集暴露,显著增强了攻击的隐蔽性与实用性。根据触发器的设计方式,后门攻击可分为静态与动态两类。静态触发器通常以固定形态嵌入样本(如特定标记或短语);而动态触发器则通过可学习的扰动策略作用于更广泛的输入空间,具有更强的泛化性和隐蔽性^[25]。

后门的植入方式主要包括数据投毒和模型篡改两种路径。通过数据投毒→后门触发,攻击者通过污染训练数据集,将含有触发器的样本与特定目标输出一同纳入训练过程,引导模型在学习过程中内化这种条件映射关系^[26]。由于生成模型训练所需数据规模庞大,来源分散,这类投毒操作极难被追踪和排查。而通过模型篡改→后门激活,攻击者则在模型训练完成后对其参数、激活路径或特定模块进行直接修改,实现后门行为的人工注入。在 LLM 中,后门一旦植入,不仅可能造成模型在特定输入下输出不当信息,泄露用户隐私或执行恶意指令;还可能通过多轮对话、多模态任务迁移等方式扩展攻击面,形成更复杂的攻击链。

3.2.5 模型窃取攻击风险

模型窃取攻击指攻击者仅在有限访问目标模型接口的情况下,通过构建输入输出对训练出功能相似的本地替代模型^[27]。该攻击多表现为查询驱动型:攻击者先设计输入样本,通过目标模型 API 获取响应,然后利用这些输入输出对训练本地模型,以近似目标模型行为。这在黑盒环境下尤为突出,攻击者无权访问模型结构和参数,仅凭输入输出数据实现功能复刻^[28]。此方式在成本和效果间具有较高性价比,攻击者无须承担大规模训练所需的资源和专业知 识,便能获得强大的 NLP 能力。模型窃取还可作为对抗样本及后门攻击提供跳板,一旦拥有替代模型,攻击者可利用白盒手段生成对抗样本或激活后门,突破防御边界。

近年来,随着 DeepSeek-R1 和 GPT-4 等高性能 LLM 的 API 的开放,其在金融、医疗、法律等专业场景的能力引发了针对性窃取研究^[29]。部分研究利用中型 Transformer 模型,通过重复调用 API 完成金融问答任务,成功训练出精度接近 GPT-3.5 的本地模型,实现绕过高额 API 调用成本的目标。

3.2.6 模型逆向攻击风险

模型逆向攻击通过模型推理阶段的输出信息,逆推出原始输入或训练数据中包含的敏感信息。这种攻击可与越狱攻击形成递进链:攻击者先逆向推断出敏感信息,再构造针对性的越狱提示,以诱导模型进一步泄露更高风险的内容。传统观点认为,语言模型编码的上下文向量仅提取语义特征,不涉及可识别个体信息。然而,Pan 等^[30]对 BERT 等 8 种通用 LLM 实施模式重建攻击发现,即便是未经保护的稠密表示,也可能泄露身份证、基因序列及患者疾病等高敏感属性。关键字推理实验显示,模型对语义特征存在记忆偏好,使攻击者可推断评论中的客户位置或医疗文本中的疾病关键词,揭示

模型文本表示可能存在过度记忆与信息泄露通道,构成潜在隐私风险。另外, Song 等^[31]的研究进一步拓展了攻击的适用范围。在白盒与黑盒场景下,通过构造辅助解码器,可有效恢复短文本嵌入中的部分关键词。即使原始输入缺乏明显结构,该方法仍能以较高精度完成逆向还原,凸显文本嵌入在隐私保护方面的脆弱性。

除此之外,基于生成式语言模型的嵌入逆向攻击更具威胁性。Li 等^[32]提出的 GEIA 方法以 GPT-2 为骨干,通过教师强制机制训练攻击模型,仅凭输入的向量表示即可生成语义一致甚至逐字重构的自然语言句子,实现从编码空间到语言空间的逆映射。Zhang 等^[33]通过 Text Revealer 技术,首次在无训练集访问权限的前提下实现了对 LLM 微调数据的近似重构。该方法通过控制 GPT-2 的隐层向量扰动,使生成文本逼近目标私有语料的分布,突破了以往攻击仅能还原关键词的局限。

3.2.7 模型越狱攻击风险

模型越狱攻击作为高度隐蔽且适应性强的对抗性提示策略,正在成为绕过模型内部安全防护和输出过滤规则的主要方式。此类攻击通过引导模型在角色认知上产生偏差,来生成原本应被拒绝的内容。例如,知名的“奶奶提示”成功诱导 ChatGPT 在“讲故事”语境下输出 Windows 10 激活密钥。

Shen 等^[34]在评估 GPT-4 对社区越狱提示集的响应效果时发现,越狱提示可将模型在敏感问题上的响应概率从 22% 提升至 56%,显示出当前 LLM 在应对语义诱导攻击方面的防御机制存在明显漏洞。

更复杂的攻击链包括:攻击者可能先通过数据投毒植入“越狱触发器”,再利用提示攻击激活触发器,最终实现信息泄露或执行恶意指令。此外,针对新一代强化对齐模型,攻击者逐步采用链式上下文构造,通过多轮对话引导巧妙绕过内置限制。Li 等^[35]研究表明,单条越狱指令已难直接奏效,但通过连续三轮模拟人机交互构建“情境陷阱”(第一轮植入越狱意图,第二轮确认模型状态,第三轮诱导输出敏感信息),可显著提升攻击成功率。这种基于上下文状态演化的“语义催眠”攻击,体现了越狱手段在系统性与策略性上的增强。面对该复杂威胁,传统依赖关键词黑名单或固定响应拒绝的安全策略已经难以有效识别和防范,亟需更智能、动态的防御机制。

综上所述,GAI 在模型层面面临多种安全风险。这些风险不仅挑战了模型的鲁棒性和可控性,也对社会信任、信息安全和伦理合规构成了严峻威胁。为了更清晰地呈现各类攻击的特征、实施方式及潜在影响,表 2 对上述 7 类模型层攻击进行了系统梳理和对比,以便于后续防御策略的分析。

表 2 GAI 模型层安全风险特征的对比分析

Table 2 Comparative analysis of security risk characteristics at GAI model level

风险类型	技术机制与攻击方式	代表性方法/研究	潜在影响	防御挑战
对抗样本攻击	通过对输入数据施加微小但有策略性的扰动,诱导模型输出错误结果	FGSM、HotFlip、基于知识蒸馏的迁移攻击等	导致模型误分类或生成偏见、虚假信息,可能引发社会对立或金融操控	文本的离散性和语义敏感性导致对抗样本难检测,现有 NLP 防御效果不足
深度伪造攻击	生成模型可合成逼真假图像或音频,支持多模态融合、端到端生成和实时交互	Stable Diffusion 等开源模型、Diffusion Models 等先进生成算法	传播虚假信息,损害个人声誉,可能影响政治、金融和社会的稳定	传统检测难以应对新型生成模型,深度伪造门槛的降低导致滥用风险上升
提示攻击	恶意输入可诱导模型偏离安全策略,包括直接指令覆盖和隐式操控两类	指令覆盖型攻击、隐式操控型攻击等	诱导模型输出错误或违规内容,可能引发隐私泄露等安全风险	生成模型的泛化与语义模糊使攻击更隐蔽,系统指令与用户输入边界亦趋模糊
后门攻击	植入触发器,操控模型输出	数据投毒、模型篡改、动态后门生成	导致模型泄露敏感信息或执行恶意操作	数据投毒隐蔽性强,后门难溯源和检测
模型窃取攻击	通过输入输出采样,克隆受限 API 模型	查询驱动型攻击、垂直领域局部克隆等	攻击者可以极低成本窃取大模型能力	黑盒攻击仅靠输入输出即可训练,现有防御难挡
模型逆向攻击	模型输出可被用于逆向还原并暴露原始敏感数据	稠密表示逆向攻击、黑盒/白盒关键词恢复等	可能泄露高敏数据,加剧隐私安全风险	模型易泄露敏感信息,现有防御机制难以有效保护语义特征
模型越狱攻击	诱导角色认知偏差,突破安全过滤生成违规内容	“奶奶提示”攻击、H-CoT 攻击等	诱导模型越界输出,引发隐私与安全风险	规则的防御难以应对复杂的上下文攻击

4 GAI 安全的应对策略与防御技术

为有效应对 GAI 系统面临的安全威胁,构建多层次、系统化的安全防护体系至关重要^[36]。如图 6 所示,在数据层面,通过 DP 保护、HE 处理和异常检测等技术手段,确保数据处理全生命周期的隐私性与安全性;在模型层面,则重点针对对抗样本攻击、后门攻击等威胁,采用对抗训练、模型水印等防御技术,显著提升模型的鲁棒性和可追溯性^[37]。下文将分别从数据层与模型层两个维度,系统阐述当前 GAI 安全的前沿防御技术与实施策略。

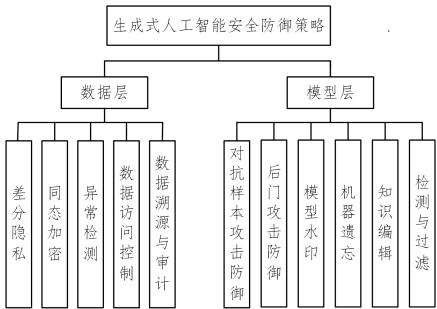


图 6 GAI 安全防御策略框架

Fig. 6 Framework of GAI security defense strategies

4.1 数据层安全防御策略

本节总结了 GAI 数据层的安全防护策略,如表 3^[38-62]所列,其中涵盖了 DP、HE、异常检测、访问控制和数据溯

源与审计五大类技术。这些方法从数据处理、访问到审计,构建了多层次防御体系,能有效应对隐私泄露、数据污染等风险。

表 3 数据层防御策略总结

Table 3 Summary of data-layer defense strategies

防御类型	作者	防御策略	优点	局限性
差分隐私	Nahid 等 ^[39]	LLM+拉普拉斯/高斯噪声注入	保留数据结构,抵御成员推理攻击	噪声可能导致模型性能下降
	Wang 等 ^[40]	低频 DP 加噪	抗滤波攻击,鲁棒性更强	计算开销增加
	Wu 等 ^[41]	自适应噪声调整	动态平衡隐私与性能	过度噪声干扰输出质量
	Lu 等 ^[42]	强化学习个性化 DP	个性化低开销	需大量交互
	Aziz 等 ^[43]	DP+HE 结合	降低 DP 的噪声注入量	实现复杂度高
同态加密	Brakerski 等 ^[44]	基于密钥转换与模转换的重线性化技术	支持任意密文计算,安全性强	计算和存储开销较大
	Guo 等 ^[45]	多等级门限秘密共享矩阵实现属性分层	细粒度访问控制,支持用户撤销	密钥管理复杂,系统实现难度高
	Matsumoto 等 ^[47]	环面全 HE	有效防止模型逆向攻击	计算复杂度高,性能瓶颈明显
	Chen 等 ^[48]	多项式编码与单次密文乘法聚合	显著提升计算效率,支持鲁棒聚合	通信成本较高,部署复杂度高
异常检测	Riaz 等 ^[49]	SMOTE+WGAN	高精度且抗模式崩溃	计算成本高,须调参
	Chen 等 ^[50]	GAN 生成+邻域比对检测异常	无监督高效,鲁棒性强	计算量大,复杂度高
	Liguori 等 ^[51]	对抗性反事实数据生成训练	提高模型对未见异常的泛化能力	生成数据质量依赖基础模型性能
	Fang 等 ^[52]	细粒度视觉异常检测方法	有效抑制自编码器异常重构能力	小规模数据集受限
	Sun 等 ^[53]	双变分 GAN	提升检测的准确性和鲁棒性	模型复杂度高,训练难度大
	Zhao 等 ^[54]	多模态条件控制生成异常样本	高泛化性,多任务支持	依赖预训练模型
数据访问控制	Cao 等 ^[55]	GPT-4V 进行多模态的异常检测	有效识别和解释零样本/单样本中的全局和细粒度语义模式	需进一步验证跨领域泛化能力
	Feretzakis 等 ^[56]	RBAC+ABAC+信任评估的动态访问控制框架	实现风险自适应的敏感信息分级披露	需要持续维护信任评估模型
	Stocks ^[57]	整数型访问控制模型	毫秒级权限验证,适合实时系统	灵活性较低,扩展性有限
	Vizgirda 等 ^[58]	基于 Solid 协议的去中心化存储与控制	用户完全掌控个人数据,增强隐私保护	存储成本较高,性能瓶颈明显
数据溯源与审计	Xie 等 ^[60]	泛化误差边界分析溯源	识别准确率高	须访问模型参数
	Belgodere 等 ^[62]	一体化合成数据可信评估框架	在高敏领域展现良好的适用性	须建立统一的溯源框架

4.1.1 差分隐私

DP 作为一种关键的隐私保护技术,在 GAI 中扮演着重要角色,尤其是在保护数据隐私的同时能确保生成任务的有效性。其基本理念是通过引入噪声,确保单一数据点的加入或删除不会显著影响模型的输出,从而降低数据泄露的风险。DP 尽管在隐私保护方面具有明显优势,但在实际应用中仍面临不少挑战^[38]。首先,噪声的引入可能导致生成模型的性能下降,特别是在文本或图像生成任务中,过度噪声会影响生成结果的连贯性和质量。其次,其噪声强度与隐私保护之间存在一定的对立关系。过强的噪声可能会导致模型输出失真,而噪声过弱则无法有效保障隐私。因此,如何平衡噪声强度与生成模型的性能,成为当前研究的关键问题。另外,DP 技术与模型层的安全防御存在潜在冲突。例如,DP 添加的噪声可能削弱模型对对抗样本的识别能力,而对抗训练中引入的扰动又可能放大 DP 带来的性能损失。这种矛盾在跨层防御设计中尤为突出,需要研究者从系统层面优化参数配置。

为了解决这些问题,近年来的研究提出了多种创新性方案。例如,Nahid 等^[39]提出的 SafeSynthDP 框架利用 LLM 结合拉普拉斯和高斯机制,在生成合成数据的过程中注入噪声,能有效保护原始数据的隐私。实验显示,在 AGNews 四分类任务中,隐私预算 $\epsilon=1$ 时,朴素贝叶斯和支持向量机的准确率分别为 77.9% 和 76.4%,仅较原始数据下降约 3% 和 10%,成员推理攻击的曲线下面积(AUC)降至 0.52,防御效果显著;在 $\epsilon=0$ 时,基于合成数据的 4 次示例上下文学习

的准确率仍达 69.8%,生成合成数据的方法显著减少了计算需求。相比传统 DP 随机梯度下降方法,该方法计算开销更低,防护效果更佳,其在 GAI 安全防护中具备实用性与优势。Wang 等^[40]提出的 LFDP 算法在 MNIST 和 CIFAR-10 数据集上,量化比较了低频 DP 与像素 DP 及生成对抗扰动等防御方法在快速梯度符号法和投影梯度下降攻击下的表现。结果显示,低频 DP 显著提升了鲁棒性半径和分类准确率,如在 MNIST 数据集上投影梯度下降攻击下的鲁棒半径从 0.0032 增至 0.6200,准确率提升 17.4%,且训练与推理时间较像素 DP 减少约 15%。这些技术方案与模型层防御存在协同效应。例如,Wu 等^[41]提出的自适应 DP 框架,根据语言项隐私概率动态调整噪声,为对抗训练提供了鲁棒性保障;Lu 等^[42]利用强化学习优化个性化 DP 联邦学习参数,兼顾了隐私保护与模型性能;Aziz 等^[43]将 DP 与 HE 结合,通过客户端及参数层匿名化处理,降低了噪声量并提升了聚合模型的准确性,简化了 HE 操作并减少了计算成本。这些创新研究推动 DP 在 GAI 中的广泛应用,既保障了隐私,又优化了生成性能。

4.1.2 同态加密

在 GAI 领域,HE 技术为数据保护提供了创新性解决方案。Brakerski 等^[44]在带误差学习假设下构建了高效的全 HE 方案。该方案引入“带密钥转换的重线性化技术”,结合模转换构建非自举层次化 HE 系统,支持在密文上直接计算任意函数。该方案可与 DP 形成互补:HE 保障计算过程的机密性,而 DP 在输出层提供可量化的隐私保证,两者结合可构

建“计算过程加密+输出结果扰动”的多层次防御体系。Guo 等^[45]提出了一种基于环上学习带误差的可撤销分层属性加密方案,通过多等级门限秘密共享实现细粒度访问控制。该结构可直接嵌入 HE 框架,利用 HE 完成密文上的策略验证与计算,实现“可撤销访问控制+密文计算”的一体化隐私保护。Zama 公司研发的 Concrete ML 框架通过自动化模型转换,将明文训练模型编译为 HE 兼容模型,支持密态推理,并可与明文环境下的对抗训练结合,实现数据层与模型层的协同防御^[46]。Matsumoto 等^[47]开发的环面全 HE 技术,通过在环面上注入连续噪声,有效防止了模型逆向攻击,提升了 GAI 系统的安全性。针对计算效率瓶颈,Chen 等^[48]提出的 AegisFL 系统融合了基于环学习带误差的近似 HE 技术与鲁棒聚合算法,通过对中间计算结果进行加密,实现多种聚合操作,在保障数据隐私的同时兼顾计算效率。该方法在 MNIST 数据集上,每轮全局训练仅需约 7s,相较于传统的帕利尔 HE 方案显著提升了模型聚合速度,通信开销与帕利尔加密方案相当,且在多种拜占庭攻击场景下,测试准确率仍能保持在接近 90% 的水平,展现出良好的实用性与鲁棒性。这些技术的创新共同构建了 GAI 的“加密计算新范式”,有效保障了数据全生命周期的隐私保护,推动了隐私保护技术在 GAI 中的广泛应用。

4.1.3 异常检测

在 GAI 领域,异常检测技术日益成为保障数据隐私和安全的重要手段。近年来,研究者提出了多种通过生成模型特性提升异常检测的鲁棒性和数据保护能力的方法。Riaz 等^[49]提出的 EO-WGAN 框架通过两阶段生成式过采样(SMOTE 初步平衡+WGAN 细化)和 Wasserstein 损失优化,在 UNSW-NB15 数据集上实现了 95.2% 的准确率,较传统 SMOTE 的准确率提升近 10%,有效解决了工业物联网中极端类别不平衡导致的异常检测难题。Chen 等^[50]提出的 GASN 模型结合 GAN 与邻域相似度分析,通过构建正常样本的潜在分布并引入邻域比对机制提升异常区分能力,在多个公开数据集上将 AUC 提升近 9.9%,明显优于传统 GAN 异常检测方法,有效增强了无监督异常检测的准确性与鲁棒性。Liguori 等^[51]提出的基于对抗性反事实生成的鲁棒异常检测方法,通过引导生成模型模拟边界样本,在 KDDCUP99 数据集上实现了 1.00 的 AUC 和 0.99 的精确率,显著优于传统对抗自编码器,提升了未见异常的泛化能力。Fang 等^[52]提出的 CKAAD 方法在 ISIC2018 和 Chest X-ray 等数据集上的表现为:仅用 1% 的粗标签,CKAAD 将 ISIC2018 的 AUC 从 87.51% 提升至 91.02%,Chest X-ray 从 85.59% 提升至 93.22%;在 5% 标签下,进一步将 AUC 提升至 92.66% 与 96.57%。在 MVTec 工业数据集上,其图像/像素级 AUC 达 99.5%/98.4%,优于 PatchCore 和 DRAEM 方法。其通过粗知识引导的对抗学习,显著提升了细粒度异常检测的精度。Sun 等^[53]提出的基于双变分 GAN 的异常检测方法,通过两个对抗网络学习正常与异常行为分布的差异,提升了检测准确性和鲁棒性。Zhao^[54]提出的 AnomalyHybrid 框架作为一种跨领域、通用的生成式异常检测方法,在无监督前提下生成多模态异常样本,在 MVTecAD 数据集上实现了图像级

97.3% 与像素级 72.9% 的平均精度指标。Cao 等^[55]评估了 GPT-4V 在 15 个多模态异常检测数据集上的性能,涵盖医疗影像、工业检测等 9 类任务,展示了大规模视觉语言模型在多模态异常检测及数据保护中的应用潜力,为 GAI 中利用异常检测技术加强数据保护提供了新的视角。

4.1.4 数据访问控制

随着 GAI 在医疗、金融、法律等敏感领域的应用,数据访问控制成为了保障数据层安全与合规的关键技术手段。传统的访问控制模型,如基于角色的访问控制(Role-Based Access Control,RBAC)和基于属性的访问控制(Attribute-Based Access Control,ABAC),在 GAI 系统中得到了进一步的发展和应用。RBAC 通过为不同角色分配权限,能有效防止未授权访问,适用于企业级 GAI 系统的精细化权限管理;ABAC 则通过评估用户、资源和环境属性,实现了动态、细粒度控制,满足复杂访问需求。

为应对日益复杂的安全需求,研究者提出多种创新访问控制机制。Feretzakis 等^[56]构建的结合用户信任评估、信息敏感性检测与输出控制的框架,利用 RBAC 和 ABAC 动态调整敏感信息披露,保障了高风险环境下的数据安全。Stocks^[57]提出的整数型访问控制模型,将安全属性编码为整数,实现了高效权限验证和数据过滤,适用于需快速响应的 GAI 系统。Vizgirda 等^[58]开发的 SocialGenPod 平台,基于 Solid 协议实现用户数据去中心化存储与访问控制,增强了用户对个人数据的掌控。

4.1.5 数据溯源与审计

随着 GAI 对大规模训练数据的依赖日益增强,数据溯源与审计机制已成为保障其数据层安全和合规性的关键技术。数据溯源可有效追踪训练数据的来源与使用情况,帮助开发者识别潜在侵权与敏感信息泄露风险。例如,Longpre 等^[59]对 1800 余个数据集进行系统审查,发现超过 70% 的数据集缺乏明确许可信息,凸显了数据透明性不足。此外,Xie 等^[60]提出的 TrainProVe 方法通过分析模型泛化误差边界,实现对模型训练数据来源的精确溯源,其在多个文本-图像生成任务中均取得极高的识别准确率。Adobe 与 Microsoft 等机构推动的 C2PA 标准已在图像生成平台中应用,能确保生成内容可验证、可追溯,从而保障内容的完整性与版权^[61]。

审计机制可评估合成数据的合规性与可信度,提升数据使用过程的透明性与责任的可追溯性。Belgodere 等^[62]提出的一体化合成数据可信评估框架,从偏差、保真度、隐私保护等维度对生成数据进行系统性量化评估,并在医疗、教育等高敏领域展示出良好的可扩展性与适用性。此外,MIT(2024)报告指出,目前多种数据审计与内容认证方法之间缺乏协同机制,亟须建立统一的溯源框架,以全面覆盖数据真实性、合法性、使用授权等关键维度^[63]。因此,推动建立以内容签名、数据文档标准、退出机制为核心的统一审计体系,是构建可追责、可信赖的 GAI 数据基础设施的关键支撑。

4.2 模型层安全防御策略

表 4 总结了 GAI 模型层的六大安全防护技术:对抗样本防御、后门攻击防御、模型水印、机器遗忘、知识编辑和检测过滤技术^[65-95]。这些方法从输入防御、参数安全、输出可控到知

识管理,构建了全链条防护体系,能有效应对对抗攻击、后门植入、版权侵犯及隐私泄露等风险。

表 4 模型层防御策略总结
Table 4 Summary of model-layer defense strategies

防御类型	作者	防御策略	优点	局限性
对抗样本防御	Gao 等 ^[65]	引入对比学习框架增强语义扰动抵抗能力	提升模型对语义扰动的鲁棒性	训练成本较高
	Wu 等 ^[66]	梯度引导的 Embedding 空间正则化,稳定语义表征	参数更新路径可控,减少语义偏移	依赖梯度计算
	Yang 等 ^[67]	双阶段掩码对抗训练	高兼容性,计算高效	对语义复杂攻击有限
后门攻击防御	Xu 等 ^[69]	对抗噪声激活,特征掩码分离	检测率高,效率高,通用性强	依赖少量干净数据,微调会略微降低准确率
	Sun 等 ^[73]	参数高效微调监测,识别适配器梯度异常	轻量化检测,适配大模型微调场景	仅适用于参数高效微调方法
	Shah 等 ^[74]	基于特征空间分布差异检测恶意模型变体	可检测微调植入的后门,适用供应链安全	需干净模型作为基准,对相似后门敏感度低
	Kirchenbauer 等 ^[76] Dathathri 等 ^[77] Ren 等 ^[78] Baldassin 等 ^[79]	红绿列表统计水印,伪随机划分词汇总表 锦标赛抽样嵌入水印,评分函数检测 语义水印算法,抵抗文本重写和同义替换 可学习的水印注意力层嵌入	生成质量无损,检测准确率高 质量无损,检测率高,计算高效 隐蔽性强,兼容语义一致性 参数占比小,抗微调能力强	对文本重写攻击鲁棒性不足 依赖部署方协作,抗编辑攻击弱 嵌入容量有限,仅支持少量信息 须修改模型结构,部署灵活性降低
模型水印	Ma 等 ^[80]	初始噪声嵌入水印+可学习嵌入网络与自适应频谱掩码	高不可见性,抗扰动能力强,保持生成图像质量	依赖生成模型结构,迁移性受限
机器遗忘	Jang 等 ^[82]	梯度上升优化目标样本,降低模型预测能力	无需全量数据重训练,计算高效	易导致非目标知识的灾难性遗忘
	Chen 等 ^[83]	轻量化遗忘层,差异化响应保留与遗忘样本	参数效率高,支持多样化删除请求	对复杂关联知识遗忘不彻底
	Zhang 等 ^[85]	负面偏好优化,仅作用于负样本区域	平衡遗忘与效用,避免性能退化	依赖精准的负面样本边界,对模糊数据效果较差
	Tian 等 ^[86]	梯度归因机制选择性淡化目标信息	精准控制遗忘范围,减少非目标知识破坏	计算复杂度高,需完整训练数据
知识编辑	Liu 等 ^[90]	基于提示词嵌入的编辑,避免直接修改权重	计算开销低,实时性强	仅适用于局部知识修正,全局一致性难保障
	Wu 等 ^[91]	选择性编辑隐私神经元,抑制隐私信息再现	精准定位隐私相关参数,降低泄露风险	须预定义隐私神经元,泛化性受限
	Venditti 等 ^[92]	调整模型关键值存储,降低隐私敏感信息泄露概率	兼容模型原有架构,操作透明	对动态新增隐私数据适应性差
检测与过滤技术	Wang 等 ^[94]	静态风险测试基准(极端主义、歧视性语言等)	透明性强,测试成本低	样本更新滞后,难应对新型攻击
	Zhang 等 ^[95]	动态检测平台,自动生成语言变异测试用例	覆盖长尾风险,适配攻击演化	生成变异可能引入评估偏差,实现复杂

4.2.1 对抗样本防御

模型层的对抗样本防御旨在提升模型对输入扰动的鲁棒性,防范对抗样本导致的误分类或敏感内容泄露等风险。对抗训练作为主流防御方法,通过引入 FGSM 和投影梯度下降等攻击算法生成的对抗样本参与训练,增强模型泛化能力^[64]。然而,该方法训练成本高,在复杂黑盒场景中效果受限,并且与其他安全威胁密切相关:模型窃取攻击获取的替代模型可为对抗攻击提供白盒环境,而成功的对抗攻击可作为后门触发器绕过传统检测。为应对这些挑战,研究者提出了基于层间鲁棒表示的学习策略,如 Gao 等^[65]提出的 SimCSE 框架通过对比学习提升了句子嵌入的语义一致性和鲁棒性。无监督条件下,其在标准 STS 任务上的平均斯皮尔曼秩相关系数为 76.3%,有监督版本提高至 81.6%,在语义扰动防御上表现优异。Wu 等^[66]提出的 SSE 方法通过嵌入正则化提升模型性能。该方法包含两种实现:SSE-Graph 利用知识图谱在 MovieLens1M 数据集上将均方根误差降低 7.6%;而 SSE-SE 无需先验知识,在 Douban 数据集上实现了 1.9%的均方根误差提升。该方法的计算效率与基线相当,在 BERT 预训练中使 SST-2 的分类准确率达到 94.5%,接近 LLM 的性能。SSE 方法通过其独特的随机嵌入共享机制,有效增强了模型对对抗样本的防御能力。Yang 等^[67]提出的 DDM 算

法通过在训练阶段随机插入掩码替代潜在的对抗干扰位,并在推理阶段动态调整可疑字符,在保持语义一致性的前提下有效中和了攻击。实验证明,该方法在多种基准数据集上均优于现有防御技术,有效增强了模型对对抗样本的防御能力。

随着生成式模型的发展,防御机制已扩展至提示级别扰动防护。防御性提示调控策略通过构建对抗提示对、多轮校验机制及强化学习优化,实现对输出内容的动态过滤。这类策略与 MU 技术存在协同效应,通过选择性遗忘恶意提示响应模式来增强防御能力。此外,基于不确定性建模的集成预测机制在应对复合攻击时展现出独特优势^[68]。检测与重构机制则通过分析激活层差异和语义一致性指标,识别对抗攻击并屏蔽潜在威胁。当前研究正从浅层扰动识别转向深层语义控制,未来须重点关注不同防御策略间的优化,以及针对复合攻击的整体防护方案设计,同时须解决防御操作与知识编辑等技术的兼容性问题,以构建更完善的 GAI 安全体系。

4.2.2 后门攻击防御

模型层的后门攻击防御策略主要聚焦于后门行为的检测、定位与消除,以防止模型在特定触发条件下输出异常响应。当前研究在后门检测机制方面已取得一定进展,主要包括基于神经激活特征的聚类分析、梯度敏感性归因、输出行为一致性验证等方法。例如,Xu 等^[69]提出的 BAN 方法利用对

抗性神经元噪声触发潜在异常激活模式,并结合神经激活特征进行聚类分析,从而有效区分干净模型与后门模型。实验表明,该方法在 CIFAR-10 与 ImageNet200 数据集上的平均检测准确率较现有方法提升约 10%,同时在运行效率上分别加快 1.37 倍与 5.11 倍,展现出高效且可扩展的检测性能。此外,部分研究引入语义扰动测试和对抗样本生成,在不破坏语义的前提下检验模型预测边界的稳定性,从而发现潜在的后门依赖关系。针对 LLM 中更具隐蔽性的触发器,通过构建上下文多轮交互环境,检测模型在特定语境中生成偏离性内容的概率,从行为层面增强后门显性暴露能力^[70]。

在后门清除与防御恢复方面,防御策略多从模型参数重构与输出行为规整入手,以切断模型对后门触发器的响应路径。其中,Sun 等^[71]提出的细粒度剪枝方法在 AISHELL 数据集上将 BadNets,PBSM 和 VSVC 攻击的攻击成功率分别降至 1.43%,22.57% 和 26.86%,同时保持了 99.4% 的干净样本准确率。该方法采用 20%~60% 渐进式剪枝并进行约 200 轮微调,训练开销低,速度与传统方法相当,在显著降低后门攻击成功率的同时,兼顾了高准确率和高效部署能力。Lin 等^[72]提出的 TSBD 框架利用少量干净样本进行退学习,通过神经元权重变化定位后门相关神经元,并采用子权重零值重初始化与梯度范数惩罚微调的两阶段策略,在 CIFAR-10 和 Tiny-ImageNet 等数据集上将多种后门攻击的 ASR 降至 1% 左右,同时保持 90% 以上的干净准确率,为仅有极少干净数据的现实场景提供了可行方案。Sun 等^[73]开发的 PEFTGuard 框架通过适配器模块梯度监测,精准识别轻量化微调下的提示注入后门。Shah 等^[74]提出的 ModelDiff 框架利用特征空间分布差异检测恶意模型变体。实验表明,该方法对模型微调过程中植入的后门具有显著的识别效果。Bai 等^[75]的研究则指出,现代后门防御体系正朝着“动态监测-静态验证”的方向发展,其中运行时检测技术可阻断激活态后门,而离线指纹比对则为模型供应链安全提供了基础保障。这些研究成果为 GAI 系统构建起多层次、渐进式的后门防御体系提供了有力支撑。

4.2.3 模型水印

随着 GAI 模型在文本、语音及图像等多模态生成任务中的广泛应用,水印技术逐渐成为保障模型输出可追溯性与防伪认证的重要手段,并在模型层安全防护体系中占据关键地位。在当前的模型水印研究中,Kirchenbauer 等^[76]提出的基于红绿列表的统计水印方法,在 C4 数据集上取得了约 98.4% 的检测灵敏度,无须访问模型内部参数,仅依赖单次前向推理即可完成验证,且生成质量保持不变。Dathathri 等^[77]提出的 SynthID-Text 框架则进一步推动了水印技术在 LLM 中的实际落地。该方法基于跳盘式抽样策略,在文本生成过程中以无失真方式嵌入水印,并结合评分函数实现高效检测。SynthID-Text 已成功部署于 Gemini 等大规模 LLM 系统,能够在不增加生成开销的前提下,帮助真实用户实时识别文本来源,并在检测性能上优于现有主流方法。

针对水印技术的隐蔽性和抗攻击性,研究者提出了多种增强方案。Ren 等^[78]开发了基于语义的水印算法,其通过词嵌入空间分析和语义值离散化处理,使水印能够抵抗文本重

写和同义替换攻击。Baldassin 等^[79]则在 Transformer 架构中引入可学习的水印注意力层,该层仅占模型参数的小部分,却能够实现水印信息的隐蔽嵌入。实验表明,即使模型经过微调,水印信息仍具备较高的保留率。Ma 等^[80]开发的 PiGW 框架则通过在初始噪声中嵌入水印,并结合可学习的嵌入网络与自适应频谱掩码,在实现高质量图像生成的同时,保持水印不可见性和抗扰动能力。这些方法为模型的安全性提供了可靠的技术支持。

4.2.4 机器遗忘

随着 GAI 模型参数量和数据依赖规模的增长,模型对训练数据的“字面记忆”能力增强,可能在推理阶段暴露个人敏感信息或版权文本,侵犯用户隐私权及“被遗忘权”。为应对这一问题,MU 作为模型层隐私防护机制被提出,用于实现对模型中已学习数据的“选择性清除”^[81]。这种机制与数据层隐私保护技术形成互补:DP 通过数据预处理防止信息泄露,而 MU 则在模型层面实现事后修正。

传统方法依赖数据预处理或重新训练,但资源消耗高且难以实际部署。因此,研究者提出了多种替代性技术路径。Jang 等^[82]提出的基于梯度上升的知识遗忘方法,在 GPT Neo 2.7B 模型上将敏感信息提取似然从 70.4% 降至 1.6%,记忆准确率从 93.4% 降至 31.0%,结构化数据遗忘效果更佳,性能下降仅约 0.4%,且通过分块顺序遗忘策略处理 128 个样本时的准确率保持在 49.8%,优于批量处理的 45.3%。Chen 等^[83]则构建轻量化的遗忘层,通过教师-学生结构对保留与遗忘样本进行差异化响应学习,仅更新小规模参数即可支持多样化删除请求。这些方法与知识编辑技术共享部分底层机制,但在实现目标和应用场景上存在显著差异。

当前的 MU 算法在提升性能与保持模型效用方面取得了一定进展,但“彻底遗忘”与“效用保持”的平衡仍具挑战。Maini 等^[84]在 TOFU 基准测试集上发现,多数梯度上升方法要么难以完全遗忘目标数据,要么导致非目标任务效能显著下降。为改善遗忘-效用权衡,Zhang 等^[85]提出负面偏好优化方法,明确优化目标仅作用于负样本区域,减少对有用知识的干扰,理论上证明了其在遗忘-效用权衡中的渐进优势。进一步地,Tian 等^[86]通过 KnowUnDo 测试集评估遗忘对非目标知识的破坏性,并据此提出 MemFlex 算法,通过精确梯度归因实现目标信息的选择性淡化。相较于模型水印技术,这类方法也涉及参数修改,但目标不同。尽管也有学者尝试从推理阶段控制记忆,如 Muresanu 等^[87]通过上下文提示构造结合典型样本聚类方式实现“冻结参数”条件下的语义遗忘,但该类方法无法真正消除模型内部的表示和记忆。Choi 等^[88]进一步揭示了遗忘过程中普遍存在的“相关性崩塌”问题,并提出 DLFD 方法,通过合成与遗忘样本的显著特征分布差异,实现对目标信息的有效擦除,同时保留任务关键相关性,在单训练轮次内即可取得优异的遗忘-效用权衡效果,为模型层遗忘机制的优化提供了新的方向。总体而言,模型层遗忘机制正由参数更新驱动走向结构级、功能级协同防护,这种演进路径与后门攻击防御技术的发展趋势高度一致。

4.2.5 知识编辑

知识编辑作为新兴隐私防护策略,可通过修改模型知识

表示,消除或修正特定信息,防止敏感数据泄露^[89]。与 MU 类似,知识编辑旨在减少模型中已学习数据的影响,但侧重点不同:MU 关注“证伪”目标数据,而知识编辑强调“证实”特定反馈,两者可协同应对模型逆向攻击。相较于传统的 MU 方法,知识编辑在计算开销和效率上具有明显优势。例如,Liu 等^[90]提出的 ECO 方法在 TOFU,WMDP 及 MMLU 等多个公开基准上展现出高效且低损的知识遗忘能力。在 TOFU 任务中,ECO 在不同遗忘比例下保持 0.99 的显著性概率值,模型效用损失低于 0.01,显著优于 NPO-KL 和 GA 方法;WMDP 测试显示,ECO 可将模型在生物、化学及网络安全任务上的准确率降至随机水平,同时对 MMLU 通用性能的影响极小。该方法通过破坏目标数据相关词嵌入实现推理阶段的隐私保护,无需微调,延迟低于 21 ms,兼具高效性与实用性。

在隐私保护领域,知识编辑的应用展现了巨大的潜力。例如,Wu 等^[91]提出的 DEPN 框架通过选择性编辑隐私神经元抑制信息再现。这种神经元级别的编辑可以与 HE 技术形成纵深防御:前者保护已泄露的信息,后者防止信息在计算过程中被泄露。Venditti 等^[92]提出的 PAE 方法在 GPT-J 模型上实现了高效的隐私保护,在 Enron 数据集上将记忆攻击成功率从 18.2%降至 11.1%~11.6%,优于 MEMIT(30.8%)和 MEND(30.7%)等基线方法;关联攻击防御的成功率达 32.4%~48.5%。PAE 单次参数更新仅需 2.3 分钟/千次编辑,其编辑效率较微调方法提升约 1250 倍,模型性能损失仅 0.33%,文本生成质量稳定,BLEU 和 METEOR 评分维持在 84 分以上。这些量化结果证实了 PAE 在防御效果、计算效率和模型性能方面具有综合优势。

然而,尽管知识编辑能够在一定程度上增强隐私保护,但如何平衡隐私保护与模型性能之间的关系仍然是一个挑战。这个挑战与前面讨论的 DP 噪声平衡问题具有相似的理论基础,都是安全与效用的权衡问题。未来的研究须建立“防御冲突量化评估”机制,通过博弈论建模分析不同防御策略间的对抗增益与协同损耗。例如,在版权保护场景中,模型水印要求底层表征稳定,知识编辑则需在高层语义空间动态调整。为缓解两者冲突,可采用分层编辑策略:在底层保留水印信息,在高层实施知识修正。这种“分层解耦”思想为破解 GAI 安全防御的“科林格里奇困境”提供了新路径。

4.2.6 检测与过滤技术

模型层的开放性使其易受恶意输入与异常输出威胁,因

此检测与过滤技术成为 GAI 安全体系中的关键防线。该类技术主要从输入检测与输出审查两个维度实现防御,旨在及时识别伪造信息、对抗提示或潜在有害内容,阻断其在生成链条中的传播。在输入端,研究者提出了多种无监督异常检测策略。例如,基于局部离群因子和密度估计的伪造样本识别方法可在缺乏标签数据的情况下判定输入是否偏离数据分布,用于检测数据注入攻击引发的模型偏移^[93]。针对诱导性提示的语义特征,已有工作引入嵌入层语义聚类 and 上下文一致性分析,以识别文本中的对抗扰动或敏感结构,保障生成语义的安全。考虑多模态系统中输入协同攻击的复杂性,一些研究还提出通过特征对齐程度分析图文一致性、语音语义同步性等指标,用于识别跨模态输入中的潜在恶意组合,提升输入端风险识别的泛化能力。

在模型输出阶段,传统的静态关键词屏蔽与规则过滤方法面临“语义绕过”与“生成变异”问题,推动研究者探索更具适应性的动态过滤机制。在内容检测方面,Wang 等^[94]构建了名为 Do-Not-Answer 的高风险指令评估静态基准,其涵盖五大风险领域和十二类危害类型,共包含 939 条高风险指令。实验结果表明,该基准能够有效评估 LLM 的安全性能。其中,LLaMA-2 表现最佳,安全率达到 99.7%;而 ChatGPT 表现相对较差,安全率为 90.9%。基于该数据集训练的 6.0×10^8 参数的 Longformer 分类器在评估效果上与 GPT-4 相当(F1 分数分别为 91.0% 和 92.6%),但在计算效率上约提升 50 倍。然而,静态基准的更新滞后性使其对抗新型攻击的防御效果从 92%降至 68%,凸显了动态评估的必要性。为突破这一局限,Zhang 等^[95]设计了 JADE 动态检测平台,通过语言变异生成复杂测试用例,挑战模型语义边界以揭示安全盲点。在生成控制方面,部分方法基于模型激活行为进行内部监测,通过比较对安全与不安全提示的响应差异识别“毒性层”,并在不破坏整体性能的前提下实施微调。例如,DINM 方法通过神经元级“手术式去毒”削弱有害生成内容;同时,神经激活轨迹分析、语义偏差评分与多轮上下文评估等机制逐步融入输出审查,构建具备行为追踪能力的“语义防火墙”,以提升模型在动态交互场景中的生成可控性与风险应对能力。

尽管现有研究提出了多种针对 GAI 安全风险的防御技术,但不同方法在实际部署中的效能存在显著差异。为此,本文选取了若干代表性防御策略,对其在公开数据集和标准攻击场景下的性能指标进行了归纳,如表 5 所列。

表 5 防御方法性能对比分析

Table 5 Comparative analysis of defense methods' performance

防御类型	文献	攻击类型	防御方法	实测数据集	防御成功率	计算开销
差分隐私	[39]	成员推理攻击	SafeSynthDP	AGNews	高	中
	[40]	FGSM PGD	LFDP	MNIST	高	中
	[41]	Canary	ADP	Wikitext-103	中	中
同态加密	[47]	数据隐私泄露	TFHE-SBC	Raspberry Pi Zero 2W	高	低
	[48]	Scaling Attack	AegisFL	MNIST	高	低
异常检测	[49]	异常检测	EO-WGAN	UNSW-NB15	高	中
	[51]	异常/对抗样本	ARN	KDDCUP99Rev	高	中
	[52]	异常图像重建	CKAAD	MVTec AD	高	中
	[54]	深度异常生成	AnomalyHybrid	MVTec AD	高	中
数据访问控制	[57]	RBAC/ABAC 绕过	IBAC	Teradata benchmark	高	低
数据溯源与审计	[59]	数据违规	数据溯源审计	DPCollection	中	中
	[60]	生成数据违规训练	TrainProVe	CIFAR10	高	中

(续表)

防御类型	文献	攻击类型	防御方法	实测数据集	防御成功率	计算开销
对抗样本防御	[65]	句子嵌入的优化	SimCSE	STS-B	高	低
	[66]	过拟合	SSE-SE	MovieLens1M	中	中
后门攻击防御	[69]	自适应攻击	BAN	CIFAR-10	高	中
	[71]	后门攻击	ASAP	AISHELL	高	中
	[72]	BadNets	TSBD	CIFAR-10	高	中
	[73]	InsertSent	PEFTGuard	IMDB	低	中
模型水印	[76]	生成式攻击	Soft Watermark	C4	高	低
	[77]	AI 生成文本检测	SynthID-Text	ELI5	高	低
	[79]	文本改写攻击	解码器层替换	OpenWebText	中	中
机器遗忘	[82]	数据提取攻击	知识遗忘	Pile	高	低
	[83]	数据提取攻击	EUL	IMDB	高	中
	[84]	隐私数据记忆攻击	Gradient Ascent	TOFU	低	低
知识编辑	[90]	危险知识提取攻击	ECO	WMDP	中	低
	[92]	训练数据提取攻击	PAE	Pile	高	低
检测与过滤	[94]	有害指令诱导攻击	BERT-like 分类	Do-Not-Answe	高	高

4.3 GAI 安全评估与应急响应体系

GAI 安全评估体系与工具平台近年来经历了从“单点测试”向“全生命周期治理”范式的跃迁。国际层面,斯坦福大学 CRFM 于 2022 年发布的 HELM 基准首次将鲁棒性、毒性、偏见与事实一致性整合为统一评分体系,对 30 余种主流大模型提供多维社会影响评估^[96]。新加坡信息通信媒体发展局与 AI Verify Foundation 于 2024 年发布《Model AI Governance Framework for Generative AI》,提出稳健性、公平性、数据治理与危险能力监控四大基线,并配套开源测试工具包,支持企业扩展垂直领域用例。国内亦取得显著进展:香港科技大学 2025 年提出的 GuidedBench 框架通过“关键实体-关键行动”评分矩阵提升攻击评估的准确性^[97];清华大学中文大模型安全评测平台集成伦理对抗题库和自动化评分引擎,可对越狱、歧视、隐私泄露等风险进行分钟级批量检测^[98];中国信通院“天秤 FlagEval”体系通过“能力-任务-指标”三维矩阵实现持续监控,并提供公共服务平台^[99]。产业界也在加速落地:蚂蚁“蚁鉴 2.0”覆盖数据投毒、后门触发及内容合规端到端检测;百度与瑞莱智慧联合推出的深度伪造检测服务,在视频平台实测中对 AI 换脸内容的实时检出率达 90% 以上。

在此基础上,GAI 安全评估工作也日益与应急响应机制深度融合,成为保障系统稳定运行的关键支柱。定期的自动化安全评估不仅有助于及时识别模型在对抗攻击、提示操控、深度伪造等方面的潜在漏洞,还可为安全防护策略的优化提供依据^[100]。例如,基于模拟攻击的测试能够量化模型在高风险情境下的稳定性,从而指导安全阈值设定与策略调整。与此同时,完备的应急响应机制成为 GAI 安全治理中的另一关键维度。其应具备自动检测、快速修复及人工干预等能力。构建完善的响应流程和事件处置预案,能够在突发安全事件中实现高效协同与快速止损,确保模型运行的连续性与用户信任的稳固性。然而,当前工具平台仍存在基准老化、模态覆盖不足、评估结果不可比等问题。针对上述问题,一是标准化演进——借鉴 NIST AI RMF 与 ISO/IEC 24713 等国际框架,构建动态基线与可扩展标准,实现跨平台、跨机构评估结果的一致性;二是红蓝对抗常态化——建立“内部红队+外部众测+漏洞赏金”的三位一体机制,强化测试强度与覆盖广度,持续暴露未知风险;三是事件驱动闭环治理——借鉴网络安全应急机制,建立 24 小时漏洞上报、72 小时补丁验证快速

响应通道。将技术基线、制度流程与实时运营深度融合,方能在创新与安全的张力中构建可信的 GAI 生态。

5 总结与展望

5.1 本文总结

GAI 作为近年来人工智能领域的关键突破之一,广泛赋能内容生成、科学研究、智能交互等多个场景,推动了语言模型、图像建模等技术体系的快速演进。然而,这种技术的复杂性与强表达能力也带来了显著的安全隐患。本文系统梳理了 GAI 在数据层和模型层面面临的主要安全挑战,探讨了当前主流的防御机制及其适用边界。

在数据层面,GAI 系统常依赖海量语料和多源异构数据进行预训练,存在隐私泄露、数据篡改与数据源不确定性等风险。本文重点分析了 DP、HE、异常检测等多种机制在应对训练数据敏感性、完整性和可信性方面的作用。这些方法在一定程度上缓解了模型“记忆效应”引发的隐私重构问题,也提升了数据源筛选与污染防御的能力。

在模型层面,对抗样本、后门注入、提示攻击以及模型逆向攻击等多类威胁成为研究重点。本文进一步评估了对抗训练、模型水印、知识编辑、机器遗忘等多项技术对提升模型鲁棒性与可控性的贡献。尤其是在预训练-指令微调范式成为主流的背景下,这些策略不仅要防范传统的白盒攻击,还须应对黑盒环境下跨模态、上下文增强等新型攻击手段。

此外,本文亦关注了近年来兴起的智能体协作攻击、多模态伪造、社会工程学操控等复杂场景中的安全隐患,这些威胁不再局限于算法漏洞,而是融合了语言认知偏差、人机交互误导与系统性协同误判,体现出 GAI 系统面向开放环境部署所面临的综合性安全挑战。

5.2 未来展望

随着 GAI 向通用智能演化,其安全治理问题日益复杂,未来研究应从以下几个方面深化与拓展。

1)多模态对抗防御框架的构建

多模态生成模型融合图像、文本、音频等异质数据,其攻击面随之扩大。未来须发展具有跨模态一致性约束的防御机制,如利用图文语义对齐进行交叉验证,采用联合注意力机制增强异常检测能力。此外,如何在保护多模态输入隐私的同时防范跨域对抗扰动,是下一步须解决的重要问题。

2)融合社会工程与人因安全的动态防控机制

GAI系统在开放交互中可能被诱导生成歧义、偏见或恶意内容,社会工程学攻击正呈现出语义驱动和心理操控双重趋势。未来的防御须引入用户意图识别与情境建模机制,结合行为分析、风险上下文评估等技术,实现对模型输入-输出路径的实时监控与响应阻断。

3)隐私攻击防御的精细化与场景化

隐私攻击正在由单轮暴露演进为上下文联动与跨模态关联建模。应探索更具适应性的DP保护策略,例如在多轮对话中动态调整噪声注入强度,或在生成路径中引入神经元级别的敏感信息屏蔽机制,以实现隐私保护与生成质量的最优平衡。

4)智能体协同攻击的建模与遏制

随着多智能体系统在任务规划、环境控制等领域的广泛应用,攻击者可能通过多轮交互和角色扮演诱导模型形成系统性误导。构建具备因果可解释性和不确定性量化能力的模型结构,如基于贝叶斯神经网络的不确定性评估与多智能体投票机制,将有助于增强系统鲁棒性与容错能力。

5)法律伦理框架与全球治理机制的完善

当前各国在AI治理政策上尚未形成共识,模型部署的跨境合规性与算法责任归属问题日益凸显。建议推动建设国际AI治理联盟,制定面向GAI的统一技术规范和伦理准则,以增强技术流通与监管协同。同时,建立透明的安全责任追溯机制,是推动GAI健康发展的制度保障。

6)全生命周期的安全评估与快速响应体系建设

面对持续演进的攻击方式,须构建覆盖训练、部署与应用的闭环安全审计系统,结合自动化测试、模拟攻击与风险评估模型,提升漏洞识别与应急响应能力。安全修复策略应具备动态适应性,并能在不破坏模型功能的前提下快速部署。

综上所述,GAI的安全问题不仅是技术难题,更是多学科交叉与全球协作的问题。构建动态、分层、可解释的安全防御体系,加强跨模态、跨任务、跨文化语境下的鲁棒性研究,将是未来保障GAI系统安全可控的核心路径。

参 考 文 献

[1] SINGH S,SINGH S,KRAUS S,et al. Characterizing generative artificial intelligence applications: Text-mining-enabled technology roadmapping [J]. Journal of Innovation and Knowledge, 2024,9(3):100531.

[2] LI Z,HUANG C,QIU W. An intrusion detection method combining variational auto-encoder and generative adversarial networks[J]. Computer Networks,2024,253:110724.

[3] CHEN J Y,XI C K,ZHENG H B,et al. A Survey on the Security of Multimodal Large Language Models [J]. Journal of Computer Science,2025,52(7):315-341.

[4] GOODFELLOW I,POUGET-ABADIE J,MIRZA M,et al. Generative adversarial nets[C]// Advances in Neural Information Processing Systems. 2014.

[5] MA Z,ZHANG Y,JIA G,et al. Efficient Diffusion Models: A Comprehensive Survey from Principles to Practices [J]. arXiv: 2410.11795,2024.

[6] HO J,JAIN A,ABBEEL P. Denoising Diffusion Probabilistic Models [J]. arXiv:2006.11239,2020.

[7] LUKOIANOV A,DE OCÁRIZ BORDE H S,GREENEWALD K,et al. Score Distillation via Reparametrized DDIM[C]// Proceedings of Advances in Neural Information Processing Systems. New York:Curran Associates Inc.,2024:26011-26044.

[8] CHEN M S,MEI S,FAN J Q,et al. Opportunities and challenges of diffusion models for generative AI[J]. National Science Review,2024,11(12):nwae348.

[9] JIANG Y,YANG Y,YIN J L,et al. A survey on security and privacy risks in large language models[J]. Journal of Computer Research and Development,2025,62(8):1979-2018.

[10] ZHANG X L,QU X Y,XIE J F,et al. Overview of Artificial Intelligence Generated Content Technologies [J/OL]. Big Data Research,2025:1-37. [2025-08-14]. <https://link.cnki.net/urlid/10.1321.G2.20250408.1439.004>.

[11] HAO Y H. Analysis and Governance of Artificial Intelligence Security Risks [J]. Journal of China Academy of Electronics and Information Technology,2020,15(6):501-505.

[12] ZHU B,MU N,JIAO J,et al. Generative AI Security:Challenges and Countermeasures [J]. arXiv:2402.12617,2024.

[13] FERRAG M A,ALWAHEDI F,BATTAH A,et al. Generative AI in Cybersecurity: A Comprehensive Review of LLM Applications and Vulnerabilities [J]. Internet of Things and Cyber-Physical Systems,2025,5:1-46.

[14] SANDMANN S,HEGSELMANN S,FUJARSKI M,et al. Benchmark evaluation of DeepSeek large language models in clinical decision-making[J]. Nature Medicine,2025,31(8):2546-2549.

[15] HU T,KYRYCHENKO Y,RATHJE S,et al. Generative language models exhibit social identity biases[J]. Nature Computational Science,2025,5:65-75.

[16] CANCELA-OUTEDA C. The EU’s AI act: A framework for collaborative governance [J]. Internet of Things, 2024, 27: 101291.

[17] HOWELLM D. Generative artificial intelligence , patient safety and healthcare quality: A review[J]. BMJ Quality & Safety, 2024,33(11):748-754.

[18] UDDIN M,IRSHAD M S,KANDHRO I A,et al. Generative AI revolution in cybersecurity:a comprehensive review of threat intelligence and operations [J]. Artificial Intelligence Review, 2025,58(8):236.

[19] KHAN N,NGUYEN T,BERMAK A,et al. Unmasking Synthetic Realities in Generative AI: A Comprehensive Review of Adversarially Robust Deepfake Detection Systems [J]. arXiv: 2507.21157,2025.

[20] YAN Z,ZHANG Y,FAN Y,et al. UCF:Uncovering Common Features for Generalizable Deepfake Detection[C]// 2023 IEEE/CVF International Conference on Computer Vision(ICCV). Piscataway,NJ:IEEE,2023:22355-22366.

[21] PEREZ E,HUANG S,SONG F,et al. Red Teaming Language Models with Language Models [J]. arXiv:2202.03286,2022.

[22] LIU Y,JIA Y,GENGR,et al. Formalizing and Benchmarking Prompt Injection Attacks and Defenses[C]//Proceedings of the

33rd USENIX Security Symposium. Philadelphia, PA: USENIX Association, 2024: 1831-1847.

[23] KANG D, LI X, STOICA I, et al. Exploiting programmatic behavior of LLMs: Chain-of-thought injections[C]// 2024 IEEE Symposium on Security and Privacy Workshops. IEEE, 2024: 132-143.

[24] LEI T, ZHANG Y, WANG X, et al. Jailbreaking black box large language models in twenty queries[C]// Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2024: 2423-2437.

[25] HUANG H, ZHAO Z Y, BACKES M, et al. Composite Backdoor Attacks Against Large Language Models[C]// Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2024. Mexico City: Association for Computational Linguistics, 2024: 1459-1472.

[26] ZHAO S, JIA M H Z, GUO Z L, et al. A Survey of Recent Backdoor Attacks and Defenses in Large Language Models[J]. arXiv: 2406. 06852, 2024.

[27] QIN Z, ZHUANG T M, ZHU G S, et al. A Survey of Security Attacks and Defense Strategies for Artificial Intelligence Models[J]. Journal of Computer Research and Development, 2024, 61(10): 2627-2648.

[28] JIAO W, HUANG J, WANG W, et al. ParroT: Translating during chat using large language models tuned with human translation and feedback[C]// Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics, 2023: 15009-15020.

[29] LIU Y, WU C, YUAN K, et al. Targeted model extraction from API-based ML services[C]// Proceedings of the 2024 IEEE Symposium on Security and Privacy. IEEE, 2024: 1843-1858.

[30] PAN X D, ZHANG M, JI S L, et al. Privacy risks of general-purpose language models[C]// Proceedings of the 41st IEEE Symposium on Security and Privacy (SP). Piscataway, NJ: IEEE, 2020: 1314-1331.

[31] SONG C Z, RAGHUNATHAN A. Information leakage in embedding models[C]// Proceedings of the 27th ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2020: 377-390.

[32] LI H, XU M S, SONG Y Q. Sentence Embedding Leaks More Information than You Expect: Generative Embedding Inversion Attack to Recover the Whole Sentence [J]. arXiv: 2305. 03010, 2023.

[33] ZHANG R, HIDANO S, KOUSHANFAR F. Text Releaver: Private Text Reconstruction via Model Inversion Attacks against Transformers [J]. arXiv: 2209. 10505, 2022.

[34] SHEN X Y, CHEN Z Y, BACKES M, et al. “Do Anything Now”: Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models [J]. arXiv: 2308. 03825, 2023.

[35] LI H, GUO D D, FAN W, et al. Multi-step Jailbreaking Privacy Attacks on ChatGPT [J]. arXiv: 2304. 05197, 2023.

[36] YANG A, YANG T A. Social Dangers of Generative Artificial Intelligence: Review and Guidelines[C]// Proceedings of the 25th Annual International Conference on Digital Government Research. 2024: 654-658.

[37] CHUA J, LI Y, YANG S Y, et al. AI Safety in Generative AI Large Language Models: A Survey [J]. arXiv: 2407. 18369, 2024.

[38] BU Z, ZHANG X, ZHA S, et al. Pre-training Differentially Private Models with Limited Public Data[C]// Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2024). Curran Associates Inc. , 2024: 94652-94683.

[39] NAHID M M H, HASAN S B. SafeSynthDP: Leveraging Large Language Models for Privacy-Preserving Synthetic Data Generation Using Differential Privacy [J]. arXiv: 2412. 20641, 2024.

[40] WANG H, XU Q, ZHANG Q H, et al. LFDP: Low-Frequency Information Fusion for Differential Privacy Robustness Enhancement [J]. Journal of Information Security, 2025, 10(1): 47-60.

[41] WU X W, GONG L, XIONG D Y. Adaptive differential privacy for language model training [C]// Proceedings of the First Workshop on Federated Learning for Natural Language Processing. Association for Computational Linguistics, 2022: 21-26.

[42] LU X Z, LIU Z H, XIAO L, et al. Reinforcement Learning-Based Personalized Differentially Private Federated Learning[J]. IEEE Transactions on Information Forensics and Security, 2025, 20: 465-477.

[43] AZIZ R, BANERJEE S, BOUZEFRANE S. Privacy preserving federated learning: a novel approach for combining differential privacy and homomorphic encryption[C]// Information Security Theory and Practice: 14th IFIP WG 11. 2 International Conference, WISTP 2024. Springer, 2024: 162-177.

[44] BRAKERSKI Z, VAIKUNTANATHAN V. Efficient Fully Homomorphic Encryption from(Standard) LWE [EB/OL]. (2011-06-26) [2025-05-21]. <https://eprint.iacr.org/2011/344>.

[45] GUO K Y, HAN Y L, WU R M. Revocable Hierarchical Attribute-Based Encryption Scheme Based on RLWE [J]. Information Technology and Network Security, 2021, 40(8): 9-16.

[46] DOBRAUNIG C, GRASSI L, HELMINGER L, et al. Pasta: a case for hybrid homomorphic encryption[C]// Proceedings of the 33rd USENIX Security Symposium. Berkeley: USENIX Association, 2024: 1-18.

[47] MATSUMOTO M, NOZAKI A, TAKASE H, et al. TFHE-SBC: Software Designs for Fully Homomorphic Encryption over the Torus on Single Board Computers [J]. arXiv: 2503. 02559, 2025.

[48] CHEN D, QU H, XU G. AegisFL: Efficient and Flexible Privacy-Preserving Byzantine-Robust Cross-silo Federated Learning [C]// Proceedings of the 41st International Conference on Machine Learning. PMLR, 2024: 7207-7219.

[49] RIAZ R, HAN G, SHAUKAT K, et al. A novel ensemble Wasserstein GAN framework for effective anomaly detection in industrial internet of things environments [J]. Scientific Reports, 2025, 15(1): 26786.

[50] CHEN L, JIANG H, WANG L, et al. Generative adversarial synthetic neighbors-based unsupervised anomaly detection [J]. Scientific Reports, 2025, 15: 16.

[51] LIGUORI A, RITACCO E, PISANI F S, et al. Robust anomaly

detection via adversarial counterfactual generation[J]. Knowledge and Information Systems,2024,66:7437-7468.

[52] FANG Q Q,SU Q L,LYU W X,et al. Boosting Fine-Grained Visual Anomaly Detection with Coarse-Knowledge-Aware Adversarial Learning[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2025;16532-16540.

[53] SUN H L,HUANG Y,HAN L S,et al. MTS-DVGAN: Anomaly detection in cyber-physical systems using a dual variational generative adversarial network [J]. Computers & Security, 2024,139:103570.

[54] ZHAO Y. Anomaly Hybrid:A Domain-agnostic Generative Framework for General Anomaly Detection[J]. arXiv:2504. 04340, 2025.

[55] CAO Y,XU X,SUN C,et al. Towards generic anomaly detection and understanding:large-scale visual-linguistic model(GPT-4V) takes the lead [J]. arXiv:2311. 02782,2023.

[56] FERETZAKIS G,VERYKIOS V S. Trustworthy AI:Securing sensitive data in large language models [J]. arXiv:2409. 18222, 2024.

[57] STOCKS M. IBAC mathematics and mechanics:The case for ‘integer based access control’ of data security in the age of AI and AI automation [J]. arXiv:2410. 19021,2024.

[58] VIZGIRDA V,ZHAO R,GOEL N. SocialGenPod:Privacy-friendly generative AI social web applications with decentralised personal data stores [C]// Proceedings of the ACM Web Conference 2024 Companion. New York:ACM,2024;1174-1178.

[59] LONGPRE S,MAHARI R,CHEN A,et al. The Data Provenance Initiative:A large scale audit of dataset licensing & attribution in AI [J]. arXiv:2310. 16787,2023.

[60] XIE Y,SONG J,WANG H,et al. Training data provenance verification:Did your model use synthetic data from my generative model for training? [J]. arXiv:2503. 09122,2025.

[61] Coalition for Content Provenance and Authenticity. Content Authenticity Initiative [EB/OL]. [2025-05-12]. <https://c2pa.org/>.

[62] BELGODERE B M,DOGNIN P L,IVANKAY A,et al. Auditing and generating synthetic data with controllable trust trade-offs[J]. IEEE Journal on Emerging and Selected Topics in Circuits and Systems,2023,14:773-788.

[63] MIT MEDIA LAB. Data provenance and consent in generative AI [EB/OL]. (2024-12-31) [2025-05-12]. <https://mit-genai.pubpub.org/pub/uk7op8zs>.

[64] ZHANG C Y,HU M W,LI W H,et al. Adversarial Attacks and Defenses on Text-to-Image Diffusion Models:A Survey[J]. arXiv:2407. 15861,2024.

[65] GAO T,YAO X,CHEN D. SimCSE:SimpleContrastive Learning of Sentence Embeddings [C]// Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics,2021;6894-6910.

[66] WU L,LI S,HSIEH C,et al. Stochastic shared embeddings:Data-driven regularization of embedding layers[C]// Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook,NY:Curran Associates Inc. ,2019: 24-33.

[67] YANG W L,YANG J,GUO Y,et al. Defensive Dual Masking for Robust Adversarial Defense[J]. arXiv:2412. 07078,2024.

[68] FRANCHI G,LAURENT O,LEGUERY M,et al. Make me a BNN: A simple strategy for estimating Bayesian uncertainty from pre-trained models[C]// 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). IEEE Computer Society,2024;12194-12204.

[69] XU X,LIU Z,KOFFAS S,et al. BAN:Detecting Backdoors Activated by Adversarial Neuron Noise[C]// Proceedings of the 38th Conference on Neural Information Processing Systems. Cambridge:MIT Press,2024;114348-114373.

[70] PAN X D,ZHANG M,SHENG B N,et al. Hidden trigger backdoor attack on NLP models via linguistic style manipulation [C]// Proceedings of the 31st USENIX Security Symposium. Berkeley:USENIX Association,2022;3611-3628.

[71] SUN H,ZHONG Q,QI M,et al. Towards robust speech models:Mitigating backdoor attacks via audio signal enhancement and fine-pruning techniques[J]. Mathematics,2025,13(6):984.

[72] LIN W,LIU L,WEI S,et al. Unveiling and Mitigating Backdoor Vulnerabilities based on Unlearning Weight Changes and Backdoor Activeness[C]// Proceedings of the 38th Conference on Neural Information Processing Systems. Cambridge:MIT Press, 2024;42097-42122.

[73] SUN Z,CONG T,LIU Y,et al. PEFTGuard:Detecting backdoor attacks against parameter-efficient fine-tuning [C]// Proceedings of the 2025 IEEE Symposium on Security and Privacy. IEEE,2025;1713-1731.

[74] SHAH H,PARK S M,ILYAS A,et al. ModelDiff:A framework for comparing learning algorithms[C]// Proceedings of the 39th International Conference on Machine Learning. PMLR, 2022: 19482-19500.

[75] BAI Y,XING G,WU H,et al. Backdoor attack and defense on deep learning:A survey[J]. IEEE Transactions on Computational Social Systems,2024,11(2):404-434.

[76] KIRSCHENBAUER J,GEIPING J,WEN Y,et al. A watermark for large language models [C]// Proceedings of the 40th International Conference on Machine Learning. PMLR,2023;17061-17084.

[77] DATHATHRI S,SEE A,GHAISAS S,et al. Scalable watermarking for identifying large language model outputs[J]. Nature,2024,634(8035):818-823.

[78] REN H,XU H,LIU Y,et al. A Robust Semantics-Based Watermark for Large Language Models Against Paraphrasing[C]// Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics,2023;7199-7213.

[79] BALDASSIN F B,NGUYEN H H,CHANG C C,et al. Cross-Attention Watermarking of Large Language Models[C]// Proceedings of the 2024 IEEE International Conference on Acoustics,Speech and Signal Processing. IEEE,2024;1-5.

[80] MA R,GUO M,LI Y,et al. PiGW:A plug-in generative watermarking framework [J]. arXiv:2403. 12053,2024.

[81] LI Z T,MENG X F,WANG L X,et al. A Survey on Machine Unlearning [J]. Journal of Software,2025,36(4):1637-1664.

[82] JANG J, YOON D, YANG S, et al. Knowledge unlearning for mitigating privacy risks in language models[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2023: 14389-14408.

[83] CHEN J, YANG D. Unlearn what you want to forget: Efficient unlearning for LLMs[C]//Proceedings of the 28th Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2023: 12041-12052.

[84] MAINI P, FENG Z, SCHWARZSCHILD A, et al. TOFU: A Task of Fictitious Unlearning for LLMs [J]. arXiv: 2401. 06121, 2024.

[85] ZHANG R, LIN L, BAI Y, et al. Negative Preference Optimization: From Catastrophic Collapse to Effective Unlearning [J]. arXiv: 2404. 05868, 2024.

[86] TIAN B, LIANG X, CHENG S, et al. To forget or not? Towards practical knowledge unlearning for large language models [C]//Proceedings of the 29th Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2024: 1524-1537.

[87] MURESANU A, THUDI A, ZHANG M R, et al. Unlearnable Algorithms for In-context Learning [J]. arXiv: 2402. 00751, 2024.

[88] CHOI D, NA D. Distribution-Level Feature Distancing for Machine Unlearning: Towards a Better Trade-off Between Model Utility and Forgetting [C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2025: 2536-2544.

[89] WANG M R, YAO Y Z, XI Z K, et al. Security Analysis of Large Model Content Generation Based on Knowledge Editing [J]. Journal of Computer Research and Development, 2024, 61(5): 1143-1155.

[90] LIU C Y, WANG Y, FLANIGAN J, et al. Large Language Model Unlearning via Embedding-Corrupted Prompts [J]. arXiv: 2406. 07933, 2024.

[91] WU X W, LI J Z, XU M H, et al. DEPN: Detecting and editing privacy neurons in pretrained language models[C]//Proceedings of the 28th Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2023: 2875-2886.

[92] VENDITTI D, RUZZETTI E S, XOMPERO G A, et al. Enhancing Data Privacy in Large Language Models through Private

Association Editing [J]. arXiv: 2406. 18221, 2024.

[93] LIU S, ZHENG C, WANG Y, et al. Unsupervised detection of adversarial examples with local outlier factors[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 5805-5818.

[94] WANG Y, LI H, HAN X, et al. Do-Not-Answer: A Dataset for Evaluating Safeguards in LLMs [J]. arXiv: 2308. 13387, 2023.

[95] ZHANG M, PAN X D, YANG M. JADE: A Linguistics-based Safety Evaluation Platform for Large Language Models[J]. arXiv: 2308. 07814, 2023.

[96] LIANG P, BOMMASANI R, LEE T, et al. Holistic evaluation of language models[J/OL]. Transactions on Machine Learning Research, 2023. <https://doi.org/10.48550/arXiv.2211.09110>.

[97] HUANG R, WANG X, LIZ, et al. Guided Bench: Measuring and Mitigating the Evaluation Discrepancies of In-the-wild LLM Jailbreak Methods [J]. arXiv: 2502. , 202516903.

[98] WANG Z, ZHANG G, YANG K, et al. Interactive Natural Language Processing [J]. arXiv: 2305. 13246, 2023.

[99] 中国信息通信研究院, 中国科学院计算技术研究所. 大模型治理蓝皮报告: 从规则走向实践[R]. 北京: 中国信通院, 2023-11-24.

[100] LEE K, KIM H, WHANG J J. SAIF: A comprehensive framework for evaluating the risks of generative AI in the public sector [J]. arXiv: 2501. 08814, 2025.



CHEN Quantao, born in 2001, postgraduate. His main research interests include natural language processing and network information security.



ZHANG Yangsen, born in 1962, postdoctoral fellow, professor, Ph.D supervisor, is a distinguished member of CCF (No. 16640D). His main research interest is natural language processing.

(责任编辑:柯颖)