特色专题

生成式人工智能安全:挑战、应对与未来

赵淦森[1,4],徐文枫[1,4],钱程[1,4],侯志豪[1,4],宁梦芹[2],龚越[2],孔广源[1,4],徐向民[5,6],郭佳宏[2*],马文俊[1,3*]

摘要 生成式人工智能作为当下智能技术发展的前沿技术,正深度重构各行业工作范式。随着生成式人工智能的广泛应用,其安全问题日益突出。系统性地审视当下的安全风险、已有的应对举措,以及未来的工作方向,成为目前的一个迫切的任务。为全面梳理生成式人工智能在各阶段的安全挑战与防护策略,从训练安全、推理安全及衍生安全3个方面系统地梳理了生成式人工智能在模型训练阶段、模型推理阶段、模型输出应用阶段的安全风险和相应的防御策略。针对训练安全和推理安全,进行了形式化建模,把具体的安全暴露面和相关要素进行了抽象的刻画,并在此基础上对相应的攻击方法和防御举措进行了深度剖析和说明。针对生成式人工智能的衍生安全问题,探讨在生成式人工智能输出结果的利用阶段所带来的信息虚假、社会公平与个人隐私等方面的安全威胁,并整理相应的应对举措。就未来人工智能的安全治理从多个方面进行讨论,包括可控性与可信生成机制、安全评估基准、责任机制与合规框架等。整体而言,对生成式人工智能的安全风险与安全防御举措进行了全面的调研,构建了一个覆盖生成式人工智能全生命周期的、覆盖技术和社会2个层面的生成式人工智能安全研究现状整体框架。

关键词 生成式人工智能;数据安全;算法安全;应用安全;社会伦理

1 概述

生成式人工智能(generative artificial intelligence, GAI)是一类人工智能算法,通过从训练数据中学习底层模式,以各种形式(如图像、文本和音乐)创建新内容。GAI是人工智能生成内容(AI generated content, AIGC)的核心技术之一,旨在基于用户输入或需求,以更快的速度和更低的成本辅助或替代人类创建内容[1]。与以分类、识别为主要任务的判别式人工智能不同,生成式人工智能更强调建模输入数据与生成结果之间的生成映射关系,具有表达能力强、任务通用性高、适应性广等特征。

GAI的技术体系经历了从早期的变分自编码器(variational auto-encoders, VAE)、生成对抗网络(generative adversarial network, GAN)到现代的Transformer等架构的快速演进。自ChatGPT提出以来,以预训练为基础,指令微调为辅助的训练模式来构建具

1. 华南师范大学,计算机学院,广州 510631
2. 北京师范大学哲学学院,逻辑与认知科学研究所,北京 100875
3. 华南师范大学,阿伯丁数据科学与人工智能学院,佛山 528225
4. 华南师范大学,广州市云计算安全与测评技术重点实验室,广州 510631
5. 佛山大学,佛山 528225
6. 人工智能与数字经济广东省实验室(广州),广州 510335

收稿日期:2025-05-27;修回日期:2026-04-22
基金项目:国家社科基金重大项目(19ZDA041)
作者简介:赵淦森,教授,研究方向为人工智能和信息安全,电子信箱:gzhao@m.scnu.edu.cn;郭佳宏(通信作者),教授,研究方向为人工智能逻辑、哲学逻辑和逻辑应用,电子信箱:jiahong.guo@bnu.edu.cn;马文俊(共同通信作者),副教授,研究方向为推荐系统、生理信号分析、不确定推理,电子信箱:mawenjun@scnu.edu.cn
引用格式:赵淦森,徐文枫,钱程,等. 生成式人工智能安全:挑战、应对与未来[J]. 科技导报, 2026, 44(12): 68-86; doi:10.3981/j.issn.1000-7857.2025.05.00145

有通用生成能力的大语言模型(large language models, LLM)逐步成为 GAI 主流。GAI 模型通过大规模语料预训练构建基础模型,在多任务生成、语言理解与推理方面展现出显著性能。

以 ChatGPT、Claude、ERNIE、DeepSeek 等为代表的大语言模型已被部署于智能问答、写作辅助等语言密集型的文本内容生成任务[2]。多模态生成模型,如 DALL·E、Stable Diffusion、MiniMax,展现出其在图像、视频等多元模态内容方面所具备的生成能力。GAI 在专业领域应用场景的决策任务中也逐渐体现出推理、规划与目标驱动的能力[3]。借助思维链、反事实推理与规划型提示工程等方法,GAI 已被用于法律文本分析、医疗问诊支持、金融决策辅助、科学文献生成与知识发现等专业任务。GAI 应用不断扩大的同时暴露出一系列的安全风险和挑战[4]。这些安全风险加剧了用户对模型输出内容的合规性、模型的可信性的担忧,同时对现有的安全治理体系与监管机制提出新挑战[5]。整体而言,GAI 的安全范畴涵盖模型训练安全、模型推理安全及模型应用所带来的衍生安全等多个维度。

1.1 安全挑战与研究动机

GAI 的全生命周期可以划分为以训练为主的模型构建阶段、以推理为主的模型部署应用阶段及模型输出后的结果应用阶段 3 大阶段。本文把 GAI 的安全按照相应的生命周期和阶段划分为训练安全、推理安全及衍生安全,重点整理和分析 GAI 所特有的安全问题。部分安全虽然与 GAI 相关,但其实是各不同领域均会出现的共性安全问题,如网络安全、系统安全等,这些安全问题不作为本文的整理和分析的重点。

在训练阶段,GAI 的构建依赖于大规模语料与复杂的训练算法,由此不可避免要面对数据层面与算法层面的安全风险。大规模训练语料的质量与真实性,对模型质量和性能影响重大。如训练语料包含恶意信息、歧视性偏见、不实信息等,则可能致使模型认知产生偏差。语料有可能包含敏感或机密的信息,在特定场景下,训练过程可能导致这些信息泄露。此外,训练算法可能存在脆弱性,使得模型容易受到对抗样本、数据投毒或梯度操控等行为的攻击,从而削弱模型的可信性。

在推理阶段,GAI 可依据用户输入指令通过推理和计算生成相应内容。此阶段,其面临的主要安全威胁源自用户利用模型输入的模型操控行为。攻击者通过构造特定的模型输入,极有可能诱导模型生成误导性、虚假乃至恶意的内容。这种攻击方法不仅可以在单轮对话中实现,还可能通过多轮交互逐步实现,甚至可能突破模型的安全限制,导致模型生成非法内容。此外,生成式人工智能受训练不充分或存在逻辑谬误等因素影响,有可能生成低质量的内容。

在结果应用阶段,GAI 面临的安全问题主要集中在其生成内容的使用和传播所造成的各类安全威胁和分享。GAI 模型输出的内容中可能包含敏感或个人隐私信息,带来个人隐私泄露和滥用风险。生成式人工智能可能会被滥用,利用其输出结果构建虚假信息、操纵舆论或加剧社会偏见,进而对社会稳定和公共信任造成负面影响。此外,GAI 引发的安全问题可能关系到国家层面,包括治理能力、主权完整和国际关系等。

综上所述,GAI 的安全风险覆盖全生命周期,涉及技术与社会多个层面。针对不同阶段的安全挑战,亟需构建多层次、系统性的防护体系,整合技术手段、伦理规范与政策监管,确保 GAI 的安全可信。

1.2 预备知识

在讨论 GAI 安全之前,有必要简要回顾 GAI 的基本原理,为后续系统性梳理和分析 GAI 的安全风险和防御举措打下基础。GAI 的核心逻辑之一是机器学习中的“找规律”。机器学习让这种找规律的过程可以被计算机自动完成。GAI 更进一步学会了前文与后文的规律,问题与答案的规律,文本与图像、与声音、与视频的规律,能够根据用户输入生成新的内容。现在的 GAI 普遍采用连接主义的技术路线,也就是用神经网络模型来模仿人脑神经元的连接关系,从大量训练数据中学习其分布。该过程可以形式化描述为一个优化问题,如公式(1)所示,含义为:寻找这样一个模型参数θ,使得模型M_θ对于前文x所预测的后文概率分布$P_\theta(x)$与训练数据D的经验分布$P_D(x)$之间的差异最小化,这样的模型参数即为最优参数θ^*。

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \mathcal{L}(P_D(\cdot|x), P_\theta(\cdot|x)) \quad (1)$$

其中,损失函数$\mathcal{L}$用于度量模型预测分布与真实数据分布之间的差异。GAI 常用的损失函数因任务不同

而有所差异,如用于文本生成任务的交叉熵损失与KL散度损失分别可以表示为

$$\mathcal{L}_{CE}(P_D,P_\theta)=-\sum_y P_D(y|x)\log P_\theta(y|x) \quad (2)$$

$$\mathcal{L}_{KL}(P_D,P_\theta)=\sum_y P_D(y|x)\log\left(\frac{P_D(y|x)}{P_\theta(y|x)}\right) \quad (3)$$

而在风格迁移、图像超分辨率和图像生成以及图像生成任务中,通常在特征空间 $\phi(\cdot)$ 上采用感知损失:

$$\mathcal{L}_{perc}=\|\phi(x)-\phi(\hat{x})\|_2^2 \quad (4)$$

然后,通过反向传播,可以计算出损失函数对于模型参数的偏导数$\nabla_\theta\mathcal{L}$,从而更新模型参数θ为

$$\theta\leftarrow\theta-\eta\nabla_\theta L \quad (5)$$

式中η是学习率。当训练收敛到一个稳定参数θ^*,模型M_{θ^*}就学到了训练数据 D 的前文与后文的规律,获得了内容 Z 的生成能力,如公式(6)所示。

$$Z=M_{\theta^*}(x) \quad (6)$$

不同的参数和输入下,模型可能会生成不同内容,部分情况下可能生成不当内容。一方面,要在训练阶段增强模型自身的鲁棒性和提升训练的质量;另一方面,要在推理阶段采取防御措施。

1.3 研究框架

本文依托GAI的全生命周期,系统性地梳理和分析各阶段的安全风险和防御举措,从而构建了“训练安全—推理安全—衍生安全”的GAI安全研究工作的3层框架。图1呈现了GAI在生命周期各阶段中的主要安全挑战及对应的防御举措。具体来说,训练安全侧重于模型在构建过程中可能遭遇的安全隐患及其对应的防御策略;推理安全关注模型推理过程中输入的安全与输出的安全;衍生安全主要探讨GAI在个人、社会和国家不同维度应用场景中可能引发的安全和伦理风险。

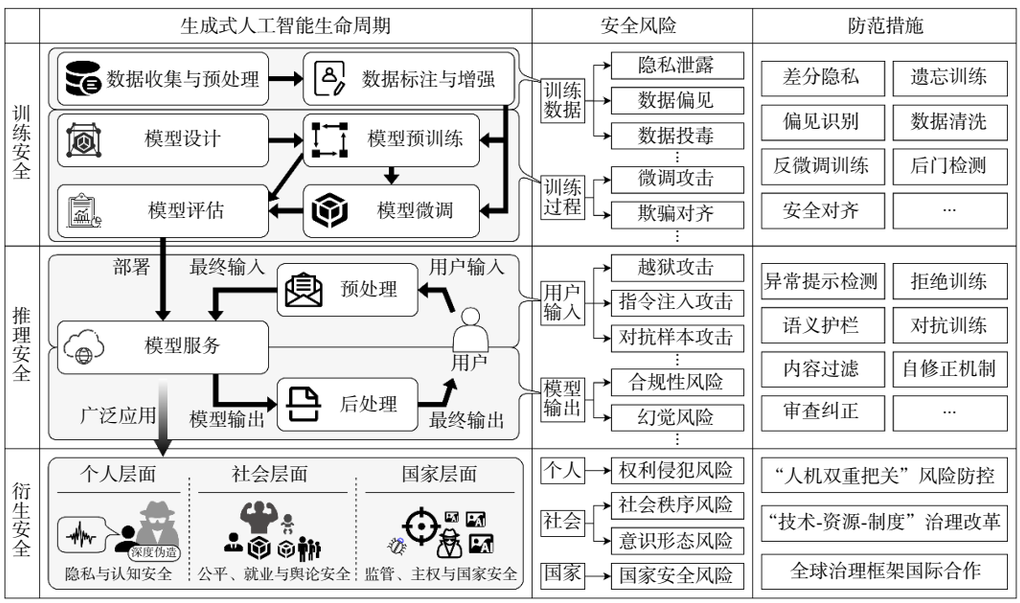

图1 研究框架总览

为确保全面性、系统性和前沿性,本文采用了系统性的文献回顾方法。研究过程包括文献检索、筛选与评估及信息综合与框架的构建。在文献检索阶段,以谷歌学术(Google Scholar)、ACM Digital Library、IEEE Xplore等国际主流学术数据库及预印本平台arXiv.org为主要文献来源,并辅以中国知网(CNKI)以覆盖国内研究现状。检索策略上,将“Generative Artificial Intelligence”“Large Language Model”等核心主题词与一系列安全关键词进行组合,其范围涵盖了从宏观的“safety”“privacy”“ethics”“alignment”等通用安全与伦理概念,到具体的“adversarial attack”“jailbreak”“prompt injection”“backdoor attack”等攻击技术,以及“hallucination”“bias”等关键现象与缓解方法。同时,结合对应的中文关键词进行检索,以保证综述的全面性。在筛选阶段,优先选择发表在高水平国际会议和期刊的同行评议论文,并且高度关注近

3 年的最新研究,对于快速发展的领域,及时纳入有影响力的 arXiv 预印本。本文定义了 5 条选择和过滤论文的规则:(1) 论文主题必须直接探讨 GAI 自身的安全、隐私、伦理或治理问题。(2) 对于某一细分领域,优先选择开创性工作、高被引论文或全面综述。(3) 排除仅将 AI 作为工具解决传统安全的文献。(4) 排除新闻、博客等非学术性文章。(5) 排除已被更新研究取代的早期文献。本文对这些文献的研究焦点进行归类统计,识别出现有的安全问题与挑战并非孤立存在,而是能够被清晰地归入模型构建(训练阶段)、模型使用(推理阶段)以及模型影响(应用衍生阶段)3 个环节。

当前,已有大量优秀的研究从不同维度对 GAI 安全进行了梳理(表 1)。Rauh 等[6]对现有安全评估方法进行了实证性梳理,指出其在模态、风险类型及上下文适应性方面存在明显的不足;Sun 等[7]提出了一个中文 LLM 的安全评估基准,涵盖了 8 种典型安全场景和 6 种指令攻击;Yao 等[8]则从宏观层面对 LLM 所带来的安全与隐私影响进行了系统分类,强调了其双重效应及内在脆弱性;Huang等[9]探讨了将传统验证与确认(verification and validation, V&V)技术应用于 LLM 的可行性,尝试为模型安全性评估提供结构化的技术路径;满超等[10]从宏观治理的视角切入,在剖析了数据、算法、应用层面风险的基础上,提出了融合法律规范、平台治理与技术监管的协同治理路径;全国网络安全标准化技术委员会[11]针对不同类型的人工智能安全风险,从技术和管理 2 方面提出防范应对措施,并针对不同角色提供安全开发应用指引。此外,数据隐私作为核心议题也得到了深入探讨。Yan 等[12]聚焦于 LLM 及智能体生命周期中的隐私保护;Liu 等[13]则综述了包括生成对抗网络和扩散模型在内的多种生成模型所面临的隐私攻击与防御策略。与现有的方法相比,本文提出的研究框架在宏观层面上概述了 GAI 在生命周期各阶段中可能面临的安全风险及现有针对性的防御措施,提供了对领域整体发展的清晰视野。

表 1 与其他安全综述论文中研究框架的对比分析

研究框架	优势	不足
Rauh等[6]	从模态、风险和上下文3个“评估差距”系统梳理了GAI现有安全评估方法	聚焦于风险评估,对防御措施的探讨相对薄弱
Yao等[8]	全面展现了LLM的正面应用与负面风险,系统梳理了多种攻击类型及其防御方法	对衍生安全的讨论较少,仅在未来工作中提及
Huang等[9]	从验证与确认的角度全面探讨了LLM的安全性和可信性问题	对衍生安全的讨论略显零散,未构建系统分析框架
Yan等[12]	聚焦隐私保护与隐私性攻击和防御,结构清晰,安全研究具有前瞻性	对衍生安全的讨论较少,仅在相关工作中提及
Liu等[13]	提供了深入的隐私攻防技术剖析,对GAI的隐私问题做了系统性总结	缺乏在衍生方面的隐私安全分析,如身份安全、认识安全等
满超等[10]	以治理/政策为主线,提供了独特的宏观治理与政策视角,与纯技术综述形成互补	对技术性安全问题的讨论不足,如后门攻击、对抗攻击等
全国网络安全标准化技术委员会[11]	作为标准/框架文件强调原则与治理,专题化概述人工智能安全风险及应对措施	对攻击和防御的技术性分析不足,如幻觉检测、隐私泄露等
Sun等[7]	聚焦中文场景,为中文LLM提供一个安全评估基准	缺乏技术性细节以及对衍生安全方面的讨论
本文框架	宏观概述GAI在全生命周期中的主要安全挑战和防御举措	宏观的概述牺牲了对单一主题的细致探讨

2 训练安全

假设训练数据集D服从数据分布$P_D(x)$,并且损失函数$\mathcal{L}$用于衡量输出生成内容的概率分布$P_\theta(x)$与$P_D(x)$的差异,那么训练过程可表示为求解使损失函数最小化的模型参数的过程。训练阶段的优化问题,模型训练过程可以形式化表示为

$$M_{\theta^*}=f_{\text{训练}}(M_\theta,D) \tag{7}$$

其中,$f_{\text{训练}}$表示训练过程,其作用是基于初始模型

M_θ 与训练数据集 D,并依据公式(1)所定义的优化目标来得到成熟模型M_{θ^*}。

从上述模型可见,决定一个 GAI 的关键要素主要为训练的数据集D、训练过程$f_{训练}$。这 2 个关键要素是训练阶段的主要安全暴露点。

在整体的流程中,训练阶段可以分为以下 5 个步骤:数据准备、预训练、微调、价值对齐和训练完成(图 2)。数据准备阶段是其中一个安全暴露位置。在数据准备阶段,训练的数据集D可能面临着隐私泄露、数据偏见、数据投毒等攻击风险。训练过程$f_{训练}$的安全风险主要出现在微调阶段或价值对齐阶段,前者主要面临微调攻击的威胁,后者主要面临着有毒反馈和奖励机制操纵的风险。

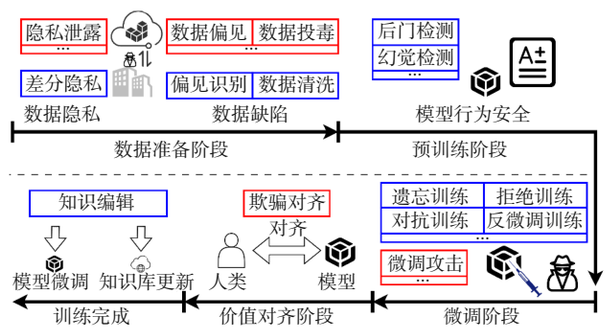

图 2 GAI 的训练安全

防御策略的实施贯穿于整个训练阶段的各个环节,涵盖了从数据准备到训练完成后的多个关键环节。差分隐私和数据清洗等数据保护技术能够很好地降低或清除数据中包含的隐私、偏见或有毒数据。在预训练阶段,后门检测、幻觉检测等防御策略能够很好地抵御数据集中残留的有毒数据的影响,确保模型的行为安全。知识遗忘、拒绝训练和反微调训练等举措能够抵御微调攻击,保护模型的有效性。价值对齐阶段可以采用一些基于规则或策略推理的方法实现模型的安全对齐。对于训练完成后仍存在的错误行为风险,知识编辑的方法能更新模型知识或借助外部知识库纠正模型的行为。

2.1 安全风险

GAI 模型面临的训练安全威胁主要体现在数据和算法层面的缺陷。这些缺陷可能是攻击者恶意攻击导致的,也可能由自身的固有缺陷引发。

针对训练数据集D的攻击,主要手段是构建存在偏见的数据、有害数据的训练数据集D',以此对模型的数据分布P_D加以引导,构建$P_{D'}$,致使模型在推理阶段针对某些方面呈现出特定偏好或偏差。训练数据集是这类型攻击的手段,其攻击对象是训练后的模型,其攻击目标是操控模型使得模型在特定条件下呈现特定的偏见等行为。常见的攻击方式包括数据偏见攻击、数据投毒攻击等。此外,针对训练数据集D的攻击类型中还包含数据泄露攻击。此类攻击主要通过多种手段,窃取训练数据集的相关内容与信息。训练数据集是这类攻击的对象,其攻击目标是窃取训练数据集的相关信息和内容。

针对训练过程$f_{训练}$的攻击,主要是通过训练过程中的有害反馈、奖励操纵等手段影响损失函数$\mathcal{L}(P_D,P_\theta)$,从而导致训练输出的模型在数据分布上出现偏差,以此对模型的行为加以引导。训练过程是这类型攻击的手段,其攻击对象是训练所得的模型,攻击目标是操控模型所对应的数据分布。具体而言,常见的攻击方式包括反向监督微调、欺骗对齐等。

2.1.1 数据安全风险

在生成式人工智能模型的训练阶段,数据不仅构成了模型能力构建的基础资源,也潜藏着多种安全风险(如隐私泄露、数据偏见、数据中毒等)。从风险来源的角度看,数据安全风险并不一定来自恶意的攻击者,也可能源于数据自身的质量缺陷。隐私泄露和数据偏见主要是由数据自身缺陷导致的。数据投毒风险主要来源于外部的攻击者。从本质上看,数据层面的攻击基本思路是通过攻击训练数据的数据分布$P_{D'}$使其偏离为错误的数据分布P_{wrong},从而使得模型学习到错误的数据分布,进而在一定程度上操控模型在推理过程中的行为。

当前模型训练实践中,隐私泄露是广泛存在的一种安全风险。隐私泄露的风险主要是由于数据未经过严格脱敏处理和模型的“记忆性”特征。模型可能在无意中复现未脱敏数据语料中的敏感片段,如个人身份信息、邮箱地址或医疗记录[14]。值得注意的是,隐私数据也可能在推理阶段被攻击者窃取,如成员推理攻击。虽然成员推理攻击通常发生在模型推理阶段,但需要在模型训练阶段进行防御。数据偏见[15]作为数据层面的自身缺陷,其问题的根源主要是训练数据集中的偏差(如数据分布不均,社会文化偏见等)。

模型在训练过程中捕捉和学习到了带有偏见的特征,导致 GAI 模型在推理过程中可能输出包含偏见的信息(如强烈的性别歧视[16])。

数据投毒是一种攻击者主动攻击以篡改训练数据分布的攻击手段,从而引导模型拟合错误的数据分布。常见的数据投毒攻击手段有标签翻转攻击、后门攻击和数据篡改攻击[17]。标签翻转攻击通过修改训练数据中部分样本的标签,使得模型在学习过程中受到误导,引发模型在分类任务中出现错误预测。后门攻击不仅篡改训练数据,还嵌入指定的触发器。当模型遇见指定的触发器时,便会输出指定的结果。数据篡改攻击不仅修改训练数据的标签或插入后门,还会直接改变数据的内容或结构,从而使得模型在学习过程中偏离真实数据分布。

2.1.2 算法安全风险

生成式模型的训练阶段可以分为预训练、微调和价值对齐 3 个阶段。其中预训练阶段依赖于大量的数据语料进行训练,其安全风险主要是数据层面的威胁。从来源角度看,微调阶段和价值对齐阶段的指令微调、反馈机制和奖励机制是容易受到攻击的暴露点。从本质上看,算法层面的攻击旨在以数据(如反向监督微调数据D_{RSFT})为手段,限制损失函数$\mathcal{L}$效果或操纵奖励机制等算法,从而使模型学习效果变差或学习到错误的知识。

微调阶段旨在通过较少的微调数据,使模型获取特定领域的知识,以增强模型的生成能力。微调阶段通常采用监督微调和强化学习微调 2 种方法[18]。其中,监督微调方法面临着有害微调的风险。攻击者通过引入恶意的微调数据(如反向监督微调[19]),从而使模型可能在特定提示下生成有害的内容。强化学习微调方法主要面临奖励操控的风险。奖励操控旨在给模型错误的引导(如调高错误行为的奖励),从而使得模型行为偏离正确方向。

价值对齐阶段旨在为确保模型的输出内容与人类伦理、道德规范等价值观层面的一致性而进行优化。模型在对齐阶段面临着欺骗对齐[20]的风险。欺骗对齐指模型可能在训练中策略性地假装对齐(符合人类价值观)。然而,模型却在实际部署时表现出真实的偏好(不符合价值观)。欺骗性对齐主要出现在较大的模型中,意味着可能受到模型规模影响。

与数据泄露类似,生成式模型也面临着模型窃取[21]的风险。攻击者通过大量查询来尝试获取模型的理论边界和数据分布特征,从而获取模型参数、隐藏的特性或者获得与被攻击模型类似效果的影子模型。与成员推理攻击相似,模型窃取攻击通常发生在模型推理阶段,但需在模型训练阶段进行防御。

2.1.3 攻击方法对比

表 2 总结了训练阶段的安全风险和攻击手段/方法,并给出了攻击目标。在数据层面,攻击者通过数据投毒主动注入后门、利用数据偏见侵蚀模型的公平性,或是通过隐私泄露反向窃取训练信息。而当威胁升级至算法层面,手段则更为精妙。微调攻击能够有针对性地篡改模型的知识,而欺骗对齐则是更难以检测的风险——AI 可能在训练中伪装顺从,部署后输出错误信息。综上所述,训练安全面临的安全威胁是利用训练数据或者训练过程,以训练数据集或者训练得到的模型作为攻击对象,最终实现窃取训练数据集或者模型的敏感信息,操控模型的数据分布从而让模型在推理阶段呈现特定的行为等。

2.2 防御措施

训练阶段的防范举措围绕着数据集D和训练过程$f_{\text{训练}}$开展。针对数据集D,主要的防范举措是保护数据集D的隐私和数据集D的数据分布,其中,隐私保护的常见方法是差分隐私和同态加密;数据分布的常见方法是数据清洗、偏见识别、后门检测、幻觉检测等。针对训练过程$f_{\text{训练}}$一种的举措是通过构建数据集

表 2 训练阶段安全风险及攻击方法对比分析

威胁层面	安全风险	攻击手段/方法	攻击目标
数据	隐私泄露	成员推理、属性推理等	从模型行为反推训练集中的敏感信息
	数据偏见	数据偏差等	模型产生系统性、非预期的不公平决策
	数据投毒	后门攻击、标签翻转等	破坏模型完整性,植入特定错误行为
算法	微调攻击	恶意样本微调等	污染或覆盖预训练知识,操控模型行为
	欺骗对齐	隐藏真实目标等	模型在训练中伪装对齐,部署后出现错误输出

$D_{拒绝}$、$D_{遗忘}$、$D_{对抗}$等,提示给模型面临攻击时的正确行为,使模型参数优化对于特定数据集(如$D_{拒绝}$)的数据分布,从而增强模型对于攻击的抵抗能力。另外一种举措是通过构建知识向量数据库,作为对模型数据分布P_D的补充,纠正P_D中可能包含的陈旧信息和虚假信息,从而增强模型生成内容的真实性和正确性。

2.2.1 数据安全防护

数据层面的隐私泄露、数据偏见和数据投毒等安全风险本质上是数据窃取或者数据分布偏差。此类安全风险可以在训练前和训练后 2 个节点进行防御。训练前,通过差分隐私、偏见识别和数据清洗等方法可以有效地减少训练数据自身缺陷,提升数据集质量。而针对数据投毒攻击,可以在训练后通过触发器重建和检测等方法进一步检测模型是否中毒。

差分隐私[22](differential privacy, DP)是一种应对隐私泄露风险的有效办法,通过在训练数据集中引入噪声等消除单个数据样本个体隐私的不可追溯性,从而避免模型输出训练数据中的隐私信息。部分研究者也尝试引入同态加密、联邦学习等机制控制训练过程中的信息暴露风险,或在生成阶段部署输出筛查模型以识别潜在敏感信息。前者在训练效率和资源消耗方面尚有较大的提升空间,后者则受限于现有识别模型对复杂语义结构的理解能力,存在漏检与误判问题。

偏见识别的方法能够有效地在训练前通过统计分析、可视化等方法识别数据集在数据分布层面的明显偏差。有研究者提出可以通过数据增强、数据合成等方法减少或者消除数据中的偏差。因偏见并非一成不变,其可能随时间发生变化(如社会价值观)。动态识别并消除数据集中隐含的偏见信息仍面临挑战。

采用可信数据源和训练前的数据清洗等传统方法能够在一定程度上缓解数据投毒攻击的威胁。然而,这种方法仅能抵御部分数据投毒攻击。针对复杂的数据投毒攻击(如后门毒药攻击),可以通过对抗性训练技术提升模型对于对抗样本(如有毒数据)的抵抗能力,避免被有毒样本影响模型效果。另一种可行的手段是使用可解释的 AI 诊断方式识别训练中引起梯度异常的有毒样本,从而抵御数据中毒攻击[18]。

2.2.2 算法安全防护

算法层面通常从模型训练中和训练后采取有效的训练方法和知识编辑等方法构建有效的防御机制来保护模型。模型训练过程中,拒绝训练、对抗性训练和幻觉检测等方法能够在训练的过程中一定程度地阻止模型学习恶意的信息和知识,降低模型在推理阶段生成虚假、恶意内容的可能,从而提高模型的鲁棒性。模型训练完成后,通过知识编辑的方法能在低成本的情况下优化模型的输出,使模型的内容更符合最新的价值观和知识体系。

针对隐私泄露问题,模型训练过程中的知识遗忘(knowledge unlearning)能够有效缓解隐私泄露风险。知识遗忘通过给模型加入遗忘层,或者引入“失忆数据集”进行微调,从而在不损害模型整体性能的前提下,有效去除不当学习内容[12]。针对 AI 幻觉问题,幻觉检测通过对模型参数施加扰动,分析其输出的波动范围,以识别模型在处理模糊或错误信息时所表现出的不确定性,从而有效监测并抑制虚假生成内容[23]。针对微调阶段的攻击,反微调训练(non-fine-tunable learning)是一种通过模拟多种潜在的微调路径,主动抑制模型在受限或高风险任务域中的学习能力,同时保持其在合法任务域内的性能表现[24]。

针对数据投毒、数据偏见等风险,对抗性训练和拒绝训练能够有效缓解这些风险。对抗训练通过将对抗样本引入训练过程,使模型在学习过程中增强对恶意扰动的识别与抵御能力。对抗样本是指在输入数据中嵌入人眼难以察觉的微小扰动,尝试诱导模型输出错误的特殊样本。拒绝训练(refusal training)通过将安全示例与带注释的有害内容纳入训练语料,引导模型学习识别并拒绝不当生成行为,强化其内容过滤能力与价值对齐能力[18]。

针对内容错误、不当模型行为、陈旧事实等问题,模型训练完成后,重新训练或者全量微调都会消耗大量的人力物力。知识编辑(knowledge editing)技术[25]能够精确地修正 GAI 模型的行为,在不影响其他无关知识的前提下完成知识更新。例如,在回答“最新一届诺贝尔化学奖获得者是谁?”问题时,采用检索增强生成(retrieval-augmented generation, RAG)方法从外部知识库或文献数据库中检索与问题相关的最新信息,再将检索结果与用户问题一并输入生成模

型,从而生成包含最新事实的回答。

2.2.3 安全防护方法对比

表 3 总结了训练阶段中可以采用的安全防御方法、优缺点,以及不同防御方法所适用的风险类型。针对数据层面,主要采用差分隐私、偏见识别和数据清洗的方法来抵御攻击。这些方法能够很好地剔除数据中隐含的偏见、虚假、隐私等信息,有力地保障了生成式人工智能的安全。但这些方法仍存在一定的缺陷,如差分隐私引入的噪声和数据清洗剔除的数据可能导致数据偏差或部分信息丢失,偏见识别依赖于先验知识识别偏见,从而具有时效性。针对算法层面,主要采用知识遗忘、幻觉检测、反微调训练、拒绝训练和知识编辑来抵御攻击。这些方法旨在避免模型学习错误知识和生成错误内容。其中,知识遗忘、反微调训练、幻觉检测和拒绝训练旨在训练和微调过程中监督模型训练和纠正模型偏差。知识编辑旨在借助外部知识修正生成内容,保证生成内容的真实性。但这些方法具有一定缺点,如知识遗忘、反微调训练和知识编辑可能对模型产生直接影响,降低模型效果。而幻觉检测和拒绝训练更依赖一定的先验知识,具有一定的时效性。因此,在实际的训练过程中,需要结合数据集和模型算法框架的实际情况,具体分析可能面临的安全风险,并选用合适的方法来应对实际的安全挑战。

表 3 安全防御方法的优缺点及适用范围

防御目标	防御方法	优点	缺点	适用范围
隐私泄露	差分隐私	数学基础、效果好	引入的噪声可能会导致生成内容质量降低	对数据脱敏,适用医疗等隐私敏感场景
	知识遗忘	无须重新训练即可删除有害知识	难以确保知识被永久遗忘,遗忘效果验证复杂	修正模型以剔除错误和隐私数据
数据偏见	偏见识别	有效识别隐含偏见	识别的时效性差,且依赖先验知识进行识别	去除训练数据中偏见信息
数据投毒	数据清洗	易操作、效果好	面对海量数据时,成本较高且可能引入偏见	对抗简单的数据投毒,基础防线
	拒绝训练	降低内容的有害性	容易被新的攻击方法绕过,需要适应新的攻击手段	可用于安全对齐,适用合法合规敏感场景
微调攻击	反微调训练	抵御微调攻击并保持性能表现	模型过于谨慎,可能导致模型通用性降低	对抗提示词注入攻击,避免模型被恶意微调
AI幻觉	幻觉检测	提升内容真实性	检测依赖定义好的事实,可能对新知识不敏感	适用于提升答案真实性
虚假内容	知识编辑	确保数据真实且时效性扩展性好	编辑的局部性难以保证,新知识可能会影响关联的旧知识,导致生成内容质量降低	适用真实性、时效性敏感任务

3 推理安全

推理阶段是 GAI 模型用其内部知识与推理能力向用户提供推理服务的过程,其本质上是一个对用户输入的三级函数映射过程。在推理阶段,公式(6)可以拆分为

$$z_{\mathrm{mo}} = M_{\theta^*}\left(u(x), y\right) \tag{8}$$

$$z_{\mathrm{fo}} = v\left(z_{\mathrm{mo}}\right) \tag{9}$$

式中,x是用户输入(如指令、上下文或提示信息),y是推理服务的外接知识库(如互联网搜索服务),z_{mo}是模型的输出,z_{fo}是推理服务的最终输出,函数M_{θ^*}是训练后的生成式人工智能模型,函数$u(\cdot)$是对用户输入的预处理,函数$v(\cdot)$是对模型输出的后处理。

在整体流程中,可以归纳为 5 个环节,即用户输入、预处理、模型推理、后处理及最终输出(图 3)。用户输入环节是接收用户的输入指令x。在该环节中,输入指令x由用户提供。因其具有开放性、可塑性和语义复杂性,成为在推理阶段主要的攻击切入点。常见的攻击手段,如越狱攻击、指令注入攻击、成员推理攻击等,是在用户输入环节中通过对输入指令x构造特定语义来显式或隐式地操控模型输出z_{mo}。

对于防御举措,除了模型推理环节中模型M_{θ^*}本身的免疫能力以外,预处理环节和后处理环节是推理

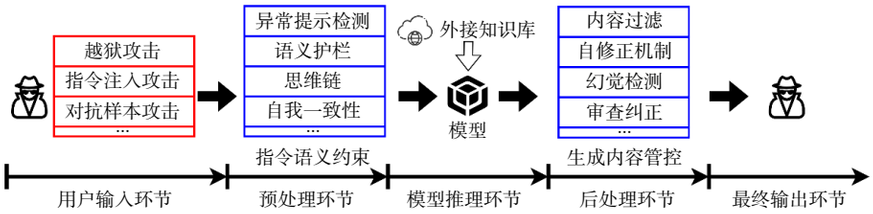

图 3 GAI 的推理安全

阶段中主要的防御面。预处理环节$u(.)$针对的是用户输入指令x,通过异常提示检测、语义护栏、思维链等技术来审查输入指令的语义。后处理环节$v(.)$针对的是模型输出内容z_{mo},通过内容过滤、自修正机制、幻觉检测等技术来检查输出内容。

3.1 攻击手段

推理阶段的攻击,通常利用自然语言的灵活多变性,以及模型理解的边界,通过特定输入诱导模型生成不当内容。可以归纳为对输入指令x添加扰动或误导性语义Δx,使得模型的输出结果偏离安全输出y',即$v(M_{\theta'}(u(x+\Delta x)))\notin y'$。具体而言,在推理过程中,GAI 主要面临来自用户侧的多种操控风险,特别是在模型输入接口高度开放的情况下。这些操控风险,如对抗样本攻击、越狱攻击和指令注入攻击,通常利用模型对自然语言的灵活解析和生成机制,通过设计特定的输入诱导模型输出不合规或潜在有害的内容。

表 4 描述了推理安全中常见的攻击手段。对抗样本攻击[26]通常通过在输入指令中加入微小的扰动来改变模型的推理结果。这些扰动虽然在表面上难以察觉,但足以引起模型生成错误的或不符合预期的输出。这类攻击利用了模型对输入分布的高度敏感性,尤其是在模型缺乏鲁棒性校准时更为有效。越狱攻击[27]则通过精心构造具有误导性的指令,绕过模型的内建安全机制,迫使模型生成违反其价值标准的内容。这类攻击不依赖于低层次扰动,而是利用自然语言的逻辑模糊性和模型对语境的过度适应性,从语义层面破坏模型的控制边界。指令注入攻击[28]通过在用户输入中嵌入恶意指令或不当语义,操控模型的响应,导致生成具有特定、通常是不合规或危险性的内容。与越狱攻击相比,指令注入攻击往往更隐蔽,攻击载荷被包装在表面合法的输入中,进而绕过简单的关键词匹配或规则过滤系统。成员推理攻击[29]和模型窃取攻击[21]的攻击方式相似,通过构建攻击样本输入模型后,观测模型的输出内容。不同的是,前者是判断样本是否用于模型训练,以达到窃取数据中成员信息的目的,后者是查看对应的响应结果,来推测模型的参数或功能,以达到复制出一个功能相似甚至完全相同模型的目的。

表 4 推理阶段安全风险及攻击方法对比分析

攻击手段	攻击过程	攻击原理
对抗样本攻击	对输入指令进行微小但有针对性的修改,诱导错误输出	利用对输入扰动的高敏感性
越狱攻击	对输入指令设计复杂或逻辑性提示,触发潜在漏洞	利用指令解析的设计缺陷
指令注入攻击	对输入指令植入特定的、隐藏的或误导性内容,利用上下文控制生成内容	利用对指令内容的无鉴别处理
成员推理攻击	访问模型的输出,推断模型的训练数据集	利用对训练数据的记忆特性
模型窃取攻击	访问模型的输出,还原模型内部参数或架构	利用输入输出间的映射规律

这些攻击方式的有效性,根源于 GAI 在结构设计与行为机制上的多重脆弱性。一方面,当前 GAI 基于条件语言建模范式,仅通过最大化下一个词元(token)的生成概率实现响应生成,缺乏对输入意图与语义边界的显式建模[30];另一方面,其开放式输入接口与上下文适应机制虽然增强了人机交互的灵活性,但也导致模型更容易受到复杂语言操控、模糊语义引导及对抗语义结构的影响。

3.2 防御措施

推理阶段的防御措施,旨在强化模型对输入指令的安全处理和输出结果的审查,从而抵御潜在的操控风险和攻击。防御策略可以从 2 个关键环节展开:首先是通过预处理函数$u(.)$对输入指令x进行严格的语义约束;其次是通过后处理函数$v(.)$对输出内容z_{mo}进

行审查,确保模型生成的内容始终符合安全标准。具体而言,预处理环节$u(.)$通过异常提示检测、语义护栏、思维链等技术,对输入指令语义进行深入筛查和约束,从源头上减少恶意操控的可能性。后处理环节$v(.)$则通过内容过滤、自修正机制和幻觉检测等技术对模型生成的内容z_{mo}进行管控,确保输出内容z_{mo}不偏离预期,并排除可能的有害信息。通过这 2 种防御机制的结合与协同作用,可有效防止通过对输入指令的扰动或误导性语义引发的模型异常行为,确保推理过程的安全性与稳定性。

3.2.1 指令语义约束

针对利用用户输入的指令进行的推理阶段的攻击,常见的指令语义约束防御措施是利用函数$u(.)$对用户输入指令进行预处理。预处理函数$u(.)$的具体实现包括对输入x进行审查和过滤,以及对输入x进行拆解和深度理解。表 5 总结了指令语义约束中常见的防御手段。

表 5 指令语义约束中常见的防御手段

防御类型	防御方法	优点	缺点	适用范围
异常提示检测	黑名单/正则表达式	实现简便,适用于显式攻击拦截	易被变体绕过,难识别语义变换	静态关键词过滤,基础防线
	语义防火墙	可识别隐蔽攻击,具备上下文理解能力	构建复杂,易受规避策略影响	对抗语义注入与任务漂移攻击
提示工程	思维链/思维树	引导结构化推理,生成稳定性与鲁棒性强	易被扰乱链条逻辑,推理成本上升	多步骤、高复杂度任务,如推理、规划
	自我一致性	多轮采样稳定性高,对单一路径依赖性低	增加生成延迟,难防深层引导型注入	精度敏感任务,如问答、代码生成

常见的防御举措是直接对输入x进行语法或者语义层面的审查和过滤。异常提示检测是常用的简单直接的输入规范化方法,这种方法主要聚焦于对输入指令中不合规内容的自动检测。这种检测关注单条指令中的敏感词、非法语义及潜在绕过语义。常见异常检测方法主要依赖预定义的黑名单与正则表达式规则对输入内容进行逐项匹配过滤,判断其是否符合预定标准[31],或者是通过距离、聚类或注意力机制判断其是否落入非法语义簇[32−33]。但这些方法缺乏对连续对话中的语义伪装或越权诱导等异常指令的预警。现有的研究提出一种语义防火墙来确保对话的安全性和合规性[34]。语义防火墙首先会将指令转化为结构化的语义图谱或动作空间表示,识别其中所蕴含的命令性行为、目标对象与语义修饰关系,然后通过依存句法分析或语义对齐网络等技术判断指令是否存在越权角色扮演或行为伪装,最后基于判定结果触发预定义的控制规则,实现输入过滤、指令修正或交互拒绝等操作[34−35]。

更深层次的防御举措是通过对输入x进行动态拆解,理解模型响应用户输入x的逻辑结构和推理路径,同时监控和审查每一个步骤的合规性。虽然异常提示检测或语义防火墙等技术能够识别出具有非法语义的指令,但是现有的攻击普遍会编织一个看似合规的指令绕过模型对输入的浅层过滤机制。为了提升 GAI 对输入意图的解析准确性与鲁棒性,提示工程可以视为输入指令后和 GAI 模型生成内容前的一道防线,通过优化模型对输入指令的解析逻辑,在内容生成前阶段建立动态安全屏障。提示工程的核心目标是引导 GAI 准确解读用户指令的语义边界。

思维链[36]是一种引导 GAI 进行逐步推理的提示工程技术。思维链以一种“导师式”的过程干预将 GAI 对指令的响应过程显式拆解为一系列结构化推理步骤,使得整个指令理解过程是可观测和可干预的。思维树[37]进一步地泛化了思维链。思维树将线性或独立的推理路径转向树状图的探索策略。它允许 GAI 逐级解析指令,形成具有逻辑层次或关联性的思维节点,这种形式能够更清晰、更拟人化地表达思考过程。自我一致性[38]结合了采样和多数投票来生成多样化的思维链。这种方式鼓励 GAI 生成不同的推理路径或对问题的不同视角。通过这些方法,指令审查机制能够实时监控模型在指令解读过程中每一步的逻辑流,能够可溯源地追踪偏误的解读过程,从而对模糊、表达隐晦或表面合法但潜在导向危险的指令具备细粒度的风险判断能力[37,39−40]。这种逐步推理方法

能够减少生成的内容中潜在的错误或误导性信息,特别是在处理复杂和多步骤问题时,模型能够清晰地解释其推理路径。

3.2.2 生成内容管控

GAI 模型在内容生成阶段的一个核心问题是生成内容的合规性。尽管 GAI 模型具备强大的语言生成能力,能够流畅地响应各类复杂指令,但是如果缺乏对法律、伦理和社会规范的内在理解,生成的内容可能触碰敏感、争议甚至违规的信息边界[41]。从基本原理上,GAI 模型本身并不会判断某句话是否“合适”,它只是依据语言模式生成概率上“合理”的文本。表 6 总结了对生成内容管控的常见防御手段。

表 6 生成内容管控中常见的防御手段

防御目标	防御手段	优点	缺点	适用范围
合规性	内容过滤	快速识别并屏蔽违规信息	可能导致误报或漏报	内容审查、合规性保障等初筛任务
	自修正机制	动态修正潜在不规范内容	算法复杂,修正延迟	高风险应用、法规遵从性要求强的场景
真实性	检索外部事实	引入外部知识库增强事实支持	依赖外部数据,可能存在滞后或不一致	事实核查、知识验证等任务
	语义熵	可量化判断生成内容的不确定性	精度受模型训练数据影响	质量控制、质量评估等任务
	引导式提示	明确推理路径,增强可控性和逻辑一致性	提示设计复杂,依赖性强	多步推理、高可解释性要求的决策任务
	链式推理	分步推理提升推导过程的透明性	长链推理易引入误差,稳定性差	复杂推理、科学推断等高精度任务

内容过滤是确保生成内容合规性的基础性手段之一。其本质是依托$z_{\mathrm{fo}}=v(z_{\mathrm{mo}})$,通过函数$v(.)$对模型输出$z_{\mathrm{mo}}$进行内容调整和修改。其目标是对模型已生成的内容进行后置筛查,识别并屏蔽其中可能涉及违法、敏感或伦理风险的信息。内容过滤与指令语义约束中提到的异常提示检测类似。前者旨在检测并屏蔽已经生成的内容中的敏感信息,而后者的目的是防止不良输入触发模型生成不当响应。内容过滤可以通过预定义专家规则或构建学习算法,自动化地剔除包含不合规内容的生成结果[42−43]。在此基础上,自修正机制进一步增强了模型在面对潜在不合规生成时的主动干预能力。该机制的核心在于引入实时验证与修正路径。当模型生成内容后,自修正机制可以通过强化学习或自监督学习等技术,与外部权威数据或内部历史记录进行比对,对生成内容即时地进行合规性验证,发现偏差后立即进行局部修正或再生成[44−45]。

GAI 模型在内容生成阶段的另一个核心问题是生成内容的真实性,即确保$z_{\mathrm{mo}}=M_{\theta^*}(u(x),y)$过程的$z_{\mathrm{mo}}$的真实性。GAI 模型依赖于大规模语料库进行训练构建起知识领域的概率分布。模型的输出z_{mo}通常是基于已知知识的概率预测,而非从真实世界的常识或实际经验中获得的信息[46]。在大模型的知识领域概率分布与真实世界不一致的时候,大模型幻觉现象可能会出现。大模型生成的内容看似有逻辑且连贯,但在事实层面却是错误或不准确的。现有的研究将这种幻觉现象分为事实性幻觉和忠实性幻觉,前者是指 GAI 模型生成的内容与可验证的现实世界事实不一致,后者是指 GAI 模型生成的内容与用户的指令或上下文不一致[47]。造成这种幻觉现象主要源于模型训练范式、语言目标设计与知识支撑机制之间的非协调性。

表 6 总结了对生成内容真实性的常见防御手段。检索外部事实和语义熵是常用的技术手段[48−51]。检索外部事实通过与可靠的知识来源进行对比,验证模型输出z_{mo}是否与外部事实一致。对于与外部事实不一致的z_{mo},一般认为模型出现幻觉。语义熵通过多轮响应,检查每一轮响应的模型输出。如果不同轮次的模型输出的一致性强,则认为输出结果z_{mo}符合真实世界。否则,一般认为模型出现了幻觉。此外,引导式提示和链式推理则是通过引导模型逐步推理,确保生成过程中的每一步都能受到逻辑和事实的一致性检查[52−54]。不同的是,前者是以用户指令为引导,而后者以模型自身推理逻辑为引导。后处理校正是在模型生成内容后对内容进行校验来审查并纠正错误内容[55]。需要提醒的是,由于 GAI 模型的底层逻辑以及

训练语料与真实世界之间的差异性,幻觉现象是GAI模型的内在缺陷。因此,幻觉现象无法被彻底解决,但是可以通过额外的策略与技术,在实际应用中尽量将幻觉的影响降到最低。

4 衍生安全

随着GAI的广泛应用,其在人类社会引发的衍生安全问题也日益凸显,这些问题不仅涉及技术层面,而且深入触及社会、文化、法律和道德等多个维度。

图4呈现了GAI的衍生安全风险所涉及的3个不同层面,即个人层面、社会层面及国家层面。个人层面重点关注的内容为个人隐私与身份安全,还有对生成式人工智能过度依赖从而导致认知削弱的风险。社会层面,主要关注点在于社会公平与数字鸿沟、经济冲击与就业替代,以及信息生态和舆论操控等方面。国家层面,主要聚焦于深层次人工智能的监管,以及国家主权与国家安全风险等问题。

4.1 个人层面:隐私与认知安全

GAI在个人层面引发的安全问题主要涉及个人基本权利的保护及人类认知削弱所带来的风险。这些问题直接关系到个体的隐私保护、心理健康和自主决策能力,构成了GAI安全风险的基础层面。

4.1.1 个人隐私与身份安全

GAI的快速发展催生了深度伪造技术(deepfake technology)[56],对个人隐私和身份安全构成了直接威胁。深度伪造利用GAI的强大生成内容,生成虚假的图像、音频和视频,捏造事实。常见的深度伪造技术包括对个人照片及视频进行高度逼真的篡改,或者利用深度伪造技术创建逼真的虚假身份特征欺骗身份认证系统。

本质上,GAI在训练过程中通过大规模真实数据学习并建模了各类分布特征,包括人脸图像、声音、影像等,从而具备了高精度、高保真的生成能力。然而,如果模型缺乏内在的价值判断机制,仅遵循概率最优原则生成输出,就可能被滥用于生成欺骗性内容或恶意工具,最终成为网络犯罪的助推器。

4.1.2 依赖与认知削弱风险

生成式人工智能的便捷性和高效性可能导致个人对其产生过度依赖,削弱人类的主体性和批判性思维能力。一方面,个体因长期依赖GAI完成思考与表达任务,导致自身认知、写作、判断等能力弱化,失去对信息的批判性思维能力,对个人专业素养形成长期负面影响;另一方面,过度依赖AI系统可能导致“去技能化”现象,即人们逐渐丧失原本具备的知识和技能。这种“去技能化”不仅影响个人能力的全面发展,还可能对社会整体的知识储备和创新能力产生负面影响。

4.2 社会层面:公平、就业与舆论安全

GAI的广泛应用可能导致就业市场变化、社会分化加剧及舆论操控等,从而在社会层面引发伦理安全问题。这些问题关系到社会公平、经济稳定和公共意识形态导向,构成了GAI安全风险的中间层面。

4.2.1 社会公平与数字鸿沟

GAI的普及虽然推动了技术民主化,但其内在的技术特性与资源分配不均却加剧了社会分层。例如,大模型的训练数据常隐含社会既有偏见(如性别、种族歧视),导致输出结果具有不公平性,而模型的黑箱特性也使得偏见难以追溯,弱势群体可能成为“算法霸凌”的受害者[57],进一步固化社会不平等。

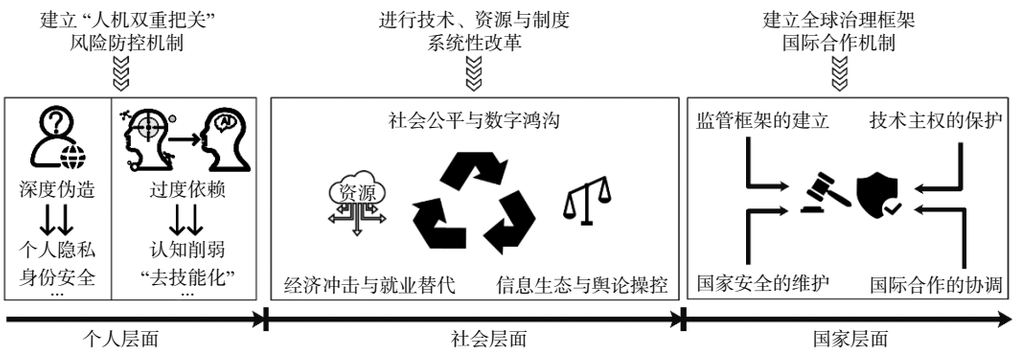

图4 GAI的衍生安全

数字鸿沟的核心在于资源分配的不平等。GAI的发展高度依赖于高性能算力、大规模训练数据与持续的能源供给。而这些资源多集中于具备技术与基础设施优势的国家、地区、平台和机构[58]。如果在制度上缺乏对资源配置的协调与激励机制,可能会在无意中扩大不同群体在技术获取、应用参与和成果分配上的差距,最终引发结构性社会公平风险。

4.2.2 经济冲击与就业替代

GAI对经济结构和行业有深远影响。在制造业,GAI技术的融入推动了生产方式的变革,可能导致对低技能劳动力的需求减少,而对技术维护、研发岗位的需求增加,企业的人力资源管理模式也会相应改变。在服务业,客服、内容生成等岗位被GAI类工具部分替代,但催生AI训练师、伦理审核员等新角色。这种行业内的调整会进而影响整个经济结构的平衡。

GAI可能会扩大就业替代的范围[59]。人工智能技术的主要目的是复制和模仿人类的智慧能力,包括学习、交互、解决问题、决策、说服和行动实施。GAI在提高生产力的同时也引发了对就业替代和劳动力需求下降的担忧。就业替代情况会逐步扩大,包括低技能与重复性岗位容易失业或被迫转向低薪临时工作,传统数据分析、文书处理、翻译等“白领”工作(如基础数据分析、文书处理)也面临挤压。另一方面,收入不平等情况可能会加剧,具体体现为GAI为一些高技能和有经验的劳动者带来了更高的收入和更多的就业机会,但低技能或中等技能的劳动者可能面临工资降低和失业的风险。

4.2.3 信息生态与舆论操控

GAI通过大规模语言模型的文本、图像及音视频生成能力,重构了信息生产与传播范式,但也对信息生态的稳定性与真实性构成严峻挑战。

从内在风险角度,GAI通过大规模语料训练和模式拟合生成内容,但技术本身无法对源头数据进行真假辨别,加上理解能力有限,容易产生“幻觉”(hallucination)现象,导致“伪知识体系”的构建成为常态。当大量未经严格验证的信息充斥网络空间时,可能引发社会性、系统性认知偏差。

从外在风险角度,GAI技术极大降低了虚假信息的生产成本和专业门槛,从而为侵害和干扰公共空间的话语权提供了新的可能性。GAI能够根据目标受众的心理特征和情感状态,定制化地生成具有特定情感影响力的内容。这种操控机制超越了传统的事实层面争议,转向对群体心理的深层干预,可能导致公共舆论的极化与理性对话空间的萎缩。

4.3 国家层面:监管、主权与国家安全风险

GAI在国家层面引发的衍生安全问题主要涉及监管框架的建立、技术主权的保护、国家安全的维护以及国际合作的协调。这些问题关系到国家治理能力、主权完整和国际关系,构成了GAI安全风险的顶层层面。

GAI对监管提出了新的挑战。GAI的训练涉及海量的数据,部分GAI平台存在滥用个人信息的行为。全球各国都在探索GAI与个人隐私保护之间的有效平衡点。各国政府纷纷出台相关法规,如欧盟的《通用数据保护条例》(GDPR)[60]、美国的《加利福尼亚消费者隐私法案》(CCPA)[61]、中国的《生成式人工智能服务管理暂行办法》等,以便更好地保护个人和组织的网络安全。同时,GAI生成内容的法律属性仍模糊——合成数据是否属于“个人数据”、生成结果的版权归属,以及由谁来承担保护这些数据的责任等问题尚无定论。

GAI的发展和应用关系到国家的技术主权和自主可控能力,技术主权面临的主要挑战包括以下2点。首先,全球AI技术发展不均衡,由于AI训练需要大量数据和计算资源,小型国家和企业可能难以建立具有竞争力的AI系统,这导致少数国家和公司控制着关键技术和基础设施,从而导致技术依赖和控制权的不平衡。其次,AI技术的快速发展使得国家在技术标准、专利和知识产权方面的竞争更加激烈,AI技术的发展已从单纯的技术竞争演变为国家主权层面的战略博弈。技术主权在GAI时代已具体化,主要涉及规则制定权、技术生态主导权和关键资源(数据、算力)控制权的多维博弈。

例如,欧盟率先制定了全面监管AI的法规《人工智能法案》(EU AI Act),该法案中构建了对技术主权的规范框架,如依据风险等级对AI系统进行分级监管[62],从而行使其规则制定权。该法案不仅是为了规范内部市场,其更深层次的战略意图是在技术竞争中捍卫和提升欧盟的自主可控能力。中美双方则围绕AI底层算法和高质量训练数据等核心资源的获取方

面存在诸多争议,GAI 时代各国在数据与算力领域加速构建自身技术体系和供应链闭环。

上述问题使得 GAI 对国家安全构成多方面的挑战,影响战略稳定和国际关系。例如,GAI 可以被用于创建虚假信息,影响公众舆论和社会稳定;这些系统也可能被用于情报收集、分析和决策,改变情报工作的性质和效果。

4.4 治理手段

生成式人工智能安全风险防控需要构建多维治理体系。在技术源头层面,建立人机协同治理机制。通过数字水印、元数据标注等构建多模态溯源体系,公开模型训练数据边界与局限性,构建虚假信息过滤验证机制,保障公共知识库安全。技术开发需要嵌入社会影响评估框架,平衡技术性能与信息生态可持续性,形成涵盖技术标准、安全评估、伦理审查的分层治理架构,对政治、医疗等高敏感领域实施算法可解释性审计和硬性约束。

社会协同治理应推进系统性改革。构建数据采集阶段的跨学科偏见审查机制,开发动态去偏算法并通过第三方公平性报告制度强化算法透明度;建立国家级算力共享网络与开源模型强制协议,完善数字税调节机制实现技术红利再分配,重点保障弱势群体数字权益。制度设计需要建立风险分层响应体系,将隐私计算、区块链技术嵌入数据全生命周期管理,强化重点行业可信计算环境建设。

全球治理维度需要建立多边协作框架。通过政府间对话机制协调技术主权与安全关切,推动行业自律标准互认及跨境应急响应协作,构建兼顾国家安全与技术共享的国际规范体系。研究表明,生成式人工智能治理需要突破传统监管范式,通过技术嵌入、制度创新与全球协作的协同演进,实现风险防控与创新发展的动态平衡。

5 展望

5.1 可控性与可信生成机制

LLM 的底层机制可以解释为基于语元关联度的形式化理论,因为关联度具有和语境相关的统计性质[63],所以 GAI 在进行下一词预测时主要依据已知知识的概率统计。从逻辑视角看,这也使得 GAI 的输入通常包含隐含假设或概率性信息,不是绝对真的信息,可能会导致非单调的结果,并不能确保其生成内容符合逻辑且真实正确。同时,在涉及认知模态的深层次推理[64-65]时,GAI 缺乏真正的认知主体性,存在语境依赖性和脆弱的一致性。尽管已有一些有效方法,如指令微调、事实校验器等,但在多轮交互和跨领域应用场景中,模型的高自由度和复杂性使得精确引导其生成结果变得困难,维持生成内容的一致性与可信性依然是难点[66-67]。面对这些挑战,应当发展更细粒度的、可解释的控制技术。例如,通过对模型的推理链进行过程监督,而非仅对最终结果进行奖惩,从而实现对生成内容的立场、风格、语气、事实来源等多维度结构化控制,提升生成过程的可控性和透明度。此外,可以借助逻辑推理的方法,将 LLM 作为“直觉提议者”,并设计外部符号引擎作为“逻辑验证者”,形成一种“提议–验证”混合架构。在这种架构中,LLM 发挥其强大的归纳与生成能力提出假设,而符号引擎则利用如可废止推理、形式论辩和反绎学习等工具,依据形式化的知识库和规则集进行严格的校验及驳斥,这不仅有助于解决信息不确定性问题,而且为黑箱透明化和可信度评估提供坚实路径问题,从而强化事实一致性保障与可信机制的集成。

5.2 安全性评估基准

GAI 的安全性评估正面临方法论碎片化与标准缺失的核心挑战。当前研究在评估维度、攻击模型、数据集选择及指标体系上缺乏统一范式[68-71],导致跨模型、跨任务的安全性结论难以横向比较,制约了防御技术的标准化发展与产业落地。构建多层次、全生命周期的标准化评估基准,成为提升 GAI 安全治理科学性的关键突破口。

1)构建分层化威胁建模体系。基于全生命周期风险特征,构建“基础层–应用层–社会层”三维威胁矩阵。基础层聚焦模型内生风险,建立对抗鲁棒性、训练数据完整性及隐私泄露概率的量化指标,例如,通过对特定攻击下的幸存率量化其鲁棒性。应用层针对多模态生成场景,开发跨模态一致性验证算法与内容合规性动态监测协议:首先,在内容安全维度,可以通过“有害内容生成率”等指标评估输出的合规性;其次,在语义安全维度,通过“属性–对象绑定准确率”等指标确保内容忠实于用户意图;最后,在规避安全维

度,通过“多模态越狱成功率”等指标衡量其对新兴攻击的防御能力。社会层引入价值对齐度评估模块,通过人类–AI 协同评估机制量化模型输出与伦理准则的偏差阈值,以判断模型在多元文化及社会争议场景下能否做出符合人类价值观的决策。

2)制定标准化验证协议。针对 GAI 安全构建覆盖“开发—测试—部署”全周期的验证流程。开发阶段实施数据谱系追溯与模型脆弱性预扫描,构建训练数据溯源路径;测试阶段部署多模态对抗攻击模拟平台,集成文本对抗扰动、图像语义篡改等攻击向量,模拟不同的威胁和攻击场景;部署阶段建立实时监控体系,通过异常检测模型动态捕获价值偏移信号与行为异常等潜在风险信号,建立与模型核心解耦的可编程护栏系统作为独立的验证与控制层,该系统通过灵活的规则引擎对模型的输入输出进行实时审查和干预,从而在不修改模型本身的前提下,敏捷地部署和更新安全策略。

3)打造动态化评估指标系统。为应对安全威胁的持续演化,针对 GAI 安全,构建动态多维评估矩阵,鲁棒性、不确定性、可解释性与价值对齐等多维度量化指标,并结合人类反馈与自动评分机制。持续跟踪 GAI 最新技术进展,挖掘新生安全挑战问题,并及时匹配对应的评估指标和方法。例如,可以建立一套能从真实世界中新发现的攻击样本和模型失败案例持续注入评估数据集的反馈闭环,以保证评估的时效性。此外,还可以将评估重心从单纯的技术指标扩展至运营级安全指标,引入“新风险遏制时间”度量安全团队从发现新漏洞到部署解决方案的响应效率。如此,实现从静态防御到动态、自适应安全治理的范式转移。

5.3 责任机制与合规框架

未来生成式人工智能安全治理的责任机制与合规框架建设需以技术价值负载性为逻辑起点,构建分层递进、多方协同的治理范式。

1)建立基于技术透明性的责任主体界定机制。针对算法黑箱引发的责任虚化问题,需要通过技术治理与制度设计的耦合实现责任穿透。具体而言,应构建算法可解释性强制披露制度,对生成式人工智能系统的决策逻辑、数据源谱系及价值嵌入路径进行全流程可审计化改造。建立“开发者–部署者–使用者”的三级连带责任体系,防止责任转嫁导致的治理真空。

2)发展动态演进的合规体系。构建“预防性合规–过程性控制–矫正性追责”的全生命周期治理路径。对模型训练数据的价值偏好进行跨学科伦理审查,建立数据集偏见指数自动监测系统。在应用部署层面实施风险分级响应机制,针对高敏场景配置嵌入式伦理审查模块,实现实时合规性校验。

3)推动伦理规范的内生化转型。合规框架需实现从被动响应向主动价值塑造的范式转换,构建“技术伦理–商业伦理–社会伦理”的三维合规体系。技术研发端应践行“生成式算法三定律”,将人类控制权、向上向善原则和可靠性要求转化为可验证的技术指标。商业模式创新需要建立伦理损益评估模型,通过数字税调节机制平衡技术创新外部性。

需要构建“政府引导–行业自律–社会监督”的协同治理生态和制度。政府层面应加快制定相关法律法规,确立算法备案、强制保险等制度工具;行业协会需开发伦理风险评估工具包,建立跨企业数据共享与风险预警机制;社会力量可通过公民技术陪审团等形式参与算法审计,形成多元共治格局。未来研究可聚焦技术透明性量化指标构建、跨域治理框架适配性测试等方向,推动责任机制从原则性宣示向操作性规范转化。

6 结论

GAI 作为由数据主体、模型训练方、算力提供方、服务运营方、终端用户、潜在攻击者与监管机构等多方耦合构成的复杂技术系统,其安全风险随着技术渗透深度的增加而呈多维扩散态势,已从单纯的技术漏洞扩展至数据伦理、社会公平及国家安全等多维度挑战。本文系统梳理了 GAI 在训练安全、推理安全及衍生安全 3 个核心阶段的安全威胁与防御策略,构建了一个覆盖全生命周期的安全研究框架,从技术机制与社会影响双重维度对 GAI 的安全挑战与防御策略进行了系统性解构。

在训练安全层面,本文揭示了数据与算法 2 个层面的安全脆弱性。数据层面的攻击通过操控训练数据分布或注入恶意样本实现模型行为的定向偏移,而算法层面的攻击则通过干扰训练过程或操纵奖励机

制削弱模型的可信性。防御策略方面,差分隐私、对抗训练与知识编辑等技术通过数据清洗、梯度修正与动态知识注入等手段,构建了从数据预处理到模型优化的多层次防护体系。然而,训练数据动态演化特性与模型规模扩展之间的安全平衡仍是待解难题。

推理安全研究聚焦输入—输出的对抗攻防机制。越狱攻击、提示注入等输入侧攻击通过语义伪装突破模型防护边界,而生成内容的合规性与真实性缺失则暴露输出侧风险。防御策略通过指令语义约束与内容生成管控的双轨制,结合思维链解析、外部知识检索与自修正机制,形成了输入过滤—过程监控—输出验证的立体防护架构。但多轮对话中的语义连贯性审查与跨模态攻击防御仍需深化。

衍生安全分析突破了传统技术安全边界,揭示了GAI对社会系统的深层重构效应。深度伪造技术对个人身份的消解、算法偏见对社会公平的侵蚀、生成内容失控对信息生态的威胁,以及技术主权博弈对国家安全的挑战,共同构成了技术社会化进程中的治理困境。应对策略建立了“技术嵌入—制度设计—国际合作”的安全治理体系,构建技术治理与社会治理的协同框架,通过数字水印、动态合规审查与全球协作机制实现风险跨域治理。

本文的学术贡献体现在3个维度:首先,通过形式化建模方法对训练与推理阶段的安全暴露面进行抽象刻画,揭示了数据投毒、梯度泄露等攻击路径的数学本质;其次,提出“技术-制度”双轮驱动的治理范式,将可解释性增强算法与合规框架建设相结合,突破传统单一技术防御的局限性;最后,构建了覆盖训练安全、推理安全与衍生安全的全生命周期安全研究工作框架,系统化梳理了相关的安全威胁、风险以及相应的应对举措。

面向未来,本文展望生成式人工智能安全研究的3个关键方向:第一,构建可信生成的理论框架,通过形式化验证与逻辑推理增强模型可控性;第二,建立统一的安全评估基准,实现威胁建模、验证协议与指标体系的标准化;第三,完善责任追溯与合规治理体系,推动技术透明性提升与伦理规范内生化。通过技术防护、制度创新与社会协同的多维联动,实现生成式人工智能的安全可信与可持续发展。

生成式人工智能的安全及其相关问题已有不少研究。本文对生成式人工智能安全领域的工作进行了全面的调研,提供了系统性理论框架与梳理,打造了覆盖训练安全、推理安全与衍生安全的全生命周期安全研究工作框架,更构建了覆盖技术演进全周期与社会影响全维度的分析范式,为后续的理论研究、行业实践与政策制定提供参考。

参考文献(References)

[1] Wang Y T, Pan Y H, Yan M, et al. A survey on ChatGPT: AI–generated contents, challenges, and solutions[J]. IEEE Open Journal of the Computer Society, 2023, 4: 280−302.

[2] 张钰莹, 云静, 刘雪颖, 等. 基于反馈的大语言模型内容与行为对齐方法综述[J]. 计算机工程与应用, 2025, 61(20): 75−104.

[3] 崔开昌, 张星辉. 生成式人工智能赋能养老金融发展的机遇、挑战与实践路径[J]. 新金融, 2024(8): 53−59.

[4] 牟奕洋, 陈涵霄, 李洪伟. 大语言模型的安全与隐私保护技术研究进展[J]. 网络空间安全科学学报, 2024, 2(1): 40−49.

[5] 赵梓羽. 生成式人工智能数据安全风险及其应对[J]. 情报资料工作, 2024, 45(2): 30−37.

[6] Rauh M, Marchal N, Manzini A, et al. Gaps in the safety evaluation of generative AI[J]. ACM Conference on AI, Ethics, and Society, 2024, 7: 1200−1217.

[7] Sun H, Zhang Z X, Deng J W, et al. Safety assessment of Chinese large language models[J/OL]. 2023: arXiv: 2304.10436.

[8] Yao Y F, Duan J H, Xu K D, et al. A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly[J]. High−Confidence Computing, 2024, 4(2): 100211.

[9] Huang X W, Ruan W J, Huang W, et al. A survey of safety and trustworthiness of large language models through the lens of verification and validation[J]. Artificial Intelligence Review, 2024, 57(7): 175.

[10] 满超, 郭永帅, 宋朝阳, 等. 生成式人工智能安全风险及协同治理研究[J]. 中国人民公安大学学报(自然科学版), 2025, 31(2): 68−73.

[11] 人工智能安全治理框架[R]. 北京: 全国网络安全标准化技术委员会, 2024.

[12] Yan B W, Li K, Xu M H, et al. On protecting the data privacy of Large Language Models (LLMs) and LLM agents: A literature review[J]. High−Confidence Computing, 2025, 5(2): 100300.

[13] Liu Y H, Huang J H, Li Y J, et al. Generative AI model privacy: A survey[J]. Artificial Intelligence Review, 2024, 58(1): 33.

[14] Gubri M, Kim S, Lee H, et al. ProPILE: Probing privacy leakage in large language models[C]//Proceedings of Advances in Neural Information Processing Systems 36.

New Orleans: Neural Information Processing Systems Foundation, Inc. , 2023: 20750−20762.

[15] Dai S H, Xu C, Xu S C, et al. Bias and unfairness in information retrieval systems: New challenges in the LLM era[C]//Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM, 2024: 6437−6447.

[16] Tang K S, Zhou W B, Zhang J, et al. GenderCARE: A comprehensive framework for assessing and reducing gender bias in large language models[C]//Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security. New York: Association for Computing Machinery, 2024: 1196−1210.

[17] Zhao P L, Zhu W Y, Jiao P F, et al. Data poisoning in deep learning: A survey[J/OL]. arXiv, 2025: 2503.22759.

[18] Wang K, Zhang G, Zhou Z, et al. A comprehensive survey in LLM(−agent) full stack safety: Data, training and deployment[J]. arXiv, 2025: 2504.15585.

[19] Yi J W, Ye R, Chen Q S, et al. On the vulnerability of safety alignment in open−access LLMs[C]//Proceedings of Findings of the Association for Computational Linguistics: ACL 2024. Kerrville: Association for Computational Linguistics 2024: 9236−9260.

[20] Greenblatt R, Denison C, Wright B, et al. Alignment faking in large language models[J/OL]. 2024: arXiv: 2412.14093.

[21] Pang K Y, Qi T, Wu C H, et al. ModelShield: Adaptive and robust watermark against model extraction attack[J]. IEEE Transactions on Information Forensics and Security, 2025, 20: 1767−1782.

[22] Yu J W, Zhou J Y, Ding Y P, et al. Textual differential privacy for context−aware reasoning with large language model[C]//Proceedings of IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC). Piscataway, NJ: IEEE, 2024: 988−997.

[23] Liu L T, Pourreza R, Panchal S, et al. Enhancing hallucination detection through noise injection[J/OL]. 2025: arXiv: 2502.03799.

[24] Deng J Y, Pang S Y, Chen Y J, et al. Sophon: Non−fine−tunable learning to restrain task transferability for pre−trained models[C]//Proceedings of IEEE Symposium on Security and Privacy (SP). Piscataway, NJ: IEEE, 2024: 2553−2571.

[25] Wang S, Zhu Y C, Liu H C, et al. Knowledge editing for large language models: A survey[J]. ACM Computing Surveys, 2025, 57(3): 1−37.

[26] Zhu K J, Wang J D, Zhou J H, et al. PromptRobust: Towards evaluating the robustness of large language models on adversarial prompts[C]//Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis. New York: Association for Computing Machinery, 2023: 57−68.

[27] Li B, Wang H H, Zhou A. Robust prompt optimization for defending language models against jailbreaking attacks[C]//Proceedings of Advances in Neural Information Processing Systems 37. Vancouver: Neural Information Processing Systems Foundation, Inc. , 2024: 40184−40211.

[28] Liu Y P, Jia Y Q, Geng R P, et al. Formalizing and benchmarking prompt injection attacks and defenses[C/OL]. Proceedings of USENIX Security Symposium, 2023: 1831−1847. https://www.usenix.org/system/files/usenixsecurity24-liu-yupei.pdf.

[29] Dziedzic A, Jia H R, Maini P, et al. LLM Dataset Inference: Did you train on my dataset? [C]//Proceedings of Advances in Neural Information Processing Systems 37. Vancouver: Neural Information Processing Systems Foundation, Inc. , 2024: 124069−124092.

[30] Hurst A, Lerer A, Goucher A P, et al. GPT−4o system card[J/OL]. arXiv, 2024: 2410.21276.

[31] Markov T, Zhang C, Agarwal S, et al. A holistic approach to undesired content detection in the real world [C]//Proceedings of the Thirty−Seventh AAAI Conference on Artificial Intelligence and Thirty−Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence. Washington, DC: AAAI Press, 2023: 15009−15018.

[32] Rahman M A, Shahriar H, Wu F, et al. Applying pre−trained multilingual BERT in embeddings for improved malicious prompt injection attacks detection[C]//Proceedings of 2nd International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings). Piscataway, NJ: IEEE, 2025: 1−7.

[33] Ayub M A, Majumdar S. Embedding−based classifiers can detect prompt injection attacks[J/OL] arXiv, 2024, 2410.22284.

[34] Rebedea T, Dinu R, Sreedhar M N, et al. *NeMo* Guardrails: A toolkit for controllable and safe LLM applications with programmable rails[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Stroudsburg, : Association for Computational Linguistics, 2023: 431−445.

[35] Wu J L, Deng J Y, Pang S Y, et al. Legilimens: Practical and unified content moderation for large language model services[C]//Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security. New York: Association for Computing Machinery, 2024: 1151−1165.

[36] Bosma M, Chi E, Ichter B, et al. Chain−of−thought prompting elicits reasoning in large language models[C]//Proceedings of Advances in Neural Information Processing Systems 35. New Orleans: Neural Information Processing Systems Foundation, Inc. , 2022: 24824−24837.

[37] Cao Y, Griffiths T, Narasimhan K, et al. Tree of thoughts:

Deliberate problem solving with large language models[C]// Proceedings of Advances in Neural Information Processing Systems 36. New Orleans: Neural Information Processing Systems Foundation, Inc. , 2023: 11809−11822.

[38] Wang X Z, Wei J, Schuurmans D, et al. Self−consistency improves chain of thought reasoning in language models[J/OL]. arXiv, 2022: 2203.11171.

[39] Sun J S, Luo Y, Gong Y Y, et al. Enhancing chain−of−thoughts prompting with iterative bootstrapping in large language models[C]//Proceedings of Findings of the Association for Computational Linguistics: NAACL 2024. Kerrville: Association for Computational Linguistics 2024: 4074−4101.

[40] Wang J N, Sun Q S, Li X, et al. Boosting language models reasoning with chain−of−knowledge prompting[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Kerrville: Association for Computational Linguistics 2024: 4958−4981.

[41] 姜毅, 杨勇, 印佳丽, 等. 大语言模型安全与隐私风险综述[J]. 计算机研究与发展, 2025, 62(8): 1979−2018.

[42] Liu J H, Shao Y Y, Zhang P, et al. Filtering discomforting recommendations with large language models[C]//Proceedings of the ACM on Web Conference 2025. New York: ACM, 2025: 3639−3650.

[43] Zeng W J, Liu Y C, Mullins R, et al. ShieldGemma: Generative AI content moderation based on Gemma[J/OL]. arXiv, 2024: 2407.21772.

[44] Yang D, Zeng L, Chen K, et al. Reinforcing Thinking through Reasoning−Enhanced Reward Models[J/OL]. arXiv, 2024: 2501.01457.

[45] Zhang Y D, Cai Y F, Zuo X Y, et al. The fusion of large language models and formal methods for trustworthy AI agents: A roadmap[J/OL]. arXiv, 2024: 2412.06512.

[46] 李煦, 朱睿, 陈小磊, 等. 视觉语言大模型的幻觉综述: 成因、评估与治理[J]. 计算机研究与发展, 2025, 62(12): 2929−2950.

[47] Huang L, Yu W J, Ma W T, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions[J]. ACM Transactions on Information Systems, 2025, 43(2): 1−55.

[48] Sahoo P, Meharia P, Ghosh A, et al. A comprehensive survey of hallucination in large language, image, video and audio foundation models[C]//Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2024. Stroudsburg: Association for Computational Linguistics, 2024: 11709−11724.

[49] Lewis P, Perez E, Piktus A, et al. Retrieval−augmented generation for knowledge−intensive NLP tasks[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc. , 2020: 9459−9474.

[50] Dhuliawala S, Komeili M, Xu J, et al. Chain−of−verification reduces hallucination in large language models[C]// Proceedings of Findings of the Association for Computational Linguistics: ACL 2024. Kerrville: Association for Computational Linguistics, 2024: 3563−3578.

[51] Peng B L, Galley M, He P C, et al. Check your facts and try again: Improving large language models with external knowledge and automated feedback[J/OL]. arXiv: 2023, 2302.12813.

[52] Cao H J, An Z W, Feng J Z, et al. A step closer to comprehensive answers: Constrained multi−stage question decomposition with large language models[J/OL]. arXiv, 2023: 2311.07491.

[53] Kang H Q, Ni J T, Yao H X. Ever: Mitigating hallucination in large language models through real−time verification and rectification[J/OL]. arXiv, 2023: 2311.09114.

[54] Zhang F J, Yu P Q, Yi B, et al. Prompt−guided internal states for hallucination detection of large language models[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Kerrville: Association for Computational Linguistics, 2025: 21806−21818.

[55] Kim J, Lee J, Chang Y, et al. Re−ex: Revising after explanation reduces the factual errors in LLM responses[J/OL]. arXiv, 2024: 2402.17097.

[56] Waqas N, Safie S I, Kadir K A, et al. DEEPFAKE image synthesis for data augmentation[J]. IEEE Access, 2022, 10: 80847−80857.

[57] O’ Neil C. Weapons of math destruction: how big data increases inequality and threatens democracy[M]. New York: Crown Publishing Group, 2016.

[58] Verdegem P. Dismantling AI capitalism: The commons as an alternative to the power concentration of Big Tech[J]. AI & Society, 2024, 39(2): 727−737.

[59] 陈小平. 人工智能伦理导引[M]. 合肥: 中国科学技术大学出版社, 2021.

[60] Lee W S, John A, Hsu H C, et al. SPChain: A smart and private blockchain−enabled framework for combining GDPR−compliant digital assets management with AI models[J]. IEEE Access, 2022, 10: 130424−130443.

[61] Stallings W. Handling of personal information and deidentified, aggregated, and pseudonymized information under the California consumer privacy act[J]. IEEE Security & Privacy, 2020, 18(1): 61−64.

[62] 丁晓东. 人工智能风险的法律规制: 以欧盟《人工智能法》为例[J]. 法律科学(西北政法大学学报), 2024, 42(5): 3−18.

[63] 陈小平. 大模型关联度预测的形式化和语义解释研究[J]. 智能系统学报, 2023, 18(4): 894−900.

[64] 罗昊轩, 郭佳宏. 描述例外的基数模态逻辑系统[J]. 逻辑

学研究, 2024, 17(5): 21−38.
[65] 宁梦芹, 罗昊轩, 郭佳宏. 动态知识−信念−意图逻辑DEKBIL及其医疗认知AI应用[J]. 逻辑学研究, 2025, 18(5): 95−119.
[66] Augenstein I, Baldwin T, Cha M, et al. Factuality challenges in the era of large language models and opportunities for fact−checking[J]. Nature Machine Intelligence, 2024, 6(8): 852−863.
[67] Zhang H Q, Song H L, Li S Y, et al. A survey of controllable text generation using transformer−based pre−trained language models[J]. ACM Computing Surveys, 2024, 56(3): 1−37.
[68] 叶文涛, 胡家齐, 王皓波, 等. 面向大语言模型安全部署的可信评估体系[J]. 计算机研究与发展, 2025, 62(7): 1668−1684.
[69] 苏艳芳, 袁静, 薛俊民. 大模型安全评估体系框架研究[C]//第39次全国计算机安全学术交流会论文集. 北京: 中国计算机学会, 2024: 107−111.
[70] 王锭, 俞怡, 姚升悦, 等. 生成式人工智能安全评估体系构建[C]//2024世界智能产业博览会人工智能安全治理主题论坛论文集. 天津: 中国网络空间安全协会, 2024: 9−13.
[71] 黑一鸣, 陈文弢, 陈杰, 等. 大模型安全风险评估与防御技术综述[J]. 中国信息安全, 2024(6): 24−27.

Security of generative artificial intelligence: Challenges, countermeasures, and future

ZHAO Gansen[1,4], XU Wenfeng[1,4], QIAN Cheng[1,4], HOU Zhihao[1,4], NING Mengqin[2], GONG Yue[2], KONG Guangyuan[1,4], XU Xiangmin[5,6], GUO Jiahong[2*], MA Wenjun[1,3*]

1. School of Computer Science, South China Normal University, Guangzhou 510631, China
2. Institute of Logic and Cognitive Science, School of Philosophy, Beijing Normal University, Beijing 100875, China
3. Aberdeen Institute of Data Science and Artificial Intelligence, South China Normal University, Foshan 528225, China
4. Key Lab on Cloud Security and Assessment technology of Guangzhou, South China Normal University, Guangzhou 510631, China
5. Foshan University, Foshan 528225, China
6. Guangdong Artificial Intelligence and Digital Economy Laboratory (Guangzhou), Guangzhou 510335, China

Abstract As a frontier technology in the current artificial intelligence era, generative artificial intelligence (GAI) is profoundly reshaping society across multiple domains. Concomitant with the proliferation of GAI applications are the emerging security challenges at both technical and social levels. To better understand the technical and social issues brought about by GAI, it is imperative to conduct a systematic and comprehensive investigation into existing security challenges, developed countermeasures, and future directions. This paper conducts a systematic survey spanning the entire lifecycle of GAI—namely, training security, inference security, and derived security—which correspond to model training, model inference, and model application, respectively. Formal models for GAI training and inference are developed to articulate security vulnerabilities, threat surfaces, and associated factors. An in−depth analysis covers typical attacks and corresponding countermeasures at both the training and inference stages. Derived security is also investigated, referring to security risks arising from GAI applications, including misinformation, social fairness concerns, individual privacy threats, and more, followed by a review of relevant countermeasures. Future directions in GAI security governance are discussed, emphasizing controllable and trustworthy generation mechanisms, security evaluation benchmarks, accountability mechanisms, and regulatory compliance frameworks. Overall, this article conducts a comprehensive investigation into the security risks and defense measures of GAI and constructs a security research framework that covers the entire lifecycle of GAI, as well as its research status at both technical and social levels.

Keywords generative artificial intelligence; training security; inference security; derivative security; social ethics

（责任编辑 王微）