智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

大语言模型安全性:分类、评估、归因、缓解、展望

黄河燕, 李思霖, 兰天伟, 邱昱力, 柳泽明, 姚嘉树, 曾理, 单赢宇, 施晓明, 郭宇航

引用本文:
黄河燕, 李思霖, 兰天伟, 等. 大语言模型安全性:分类、评估、归因、缓解、展望[J]. 智能系统学报, 2025, 20(1): 2-32.
HUANG Heyan, LI Silin, LAN Tianwei, et al. A survey on the safety of large language model: classification, evaluation, attribution, mitigation and prospect[J]. *CAAI Transactions on Intelligent Systems*, 2025, 20(1): 2-32.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202401006

DOI: 10.11992/tis.202401006
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20241212.1608.002

大语言模型安全性: 分类、评估、归因、缓解、展望

黄河燕[1], 李思霖[1], 兰天伟[1], 邱昱力[1], 柳泽明[2], 姚嘉树[1], 曾理[1], 单赢宇[1], 施晓明[3], 郭宇航[1]

(1. 北京理工大学 计算机学院, 北京 100081; 2. 北京航空航天大学 计算机学院, 北京 100191; 3. 哈尔滨工业大学 计算机学院社会计算与信息检索研究中心, 黑龙江 哈尔滨 150001)

摘 要: 大语言模型能够在多个领域及任务上给出与人类水平相当的解答, 并且在未经训练的领域和任务上展现了丰富的涌现能力。然而, 目前基于大语言模型的人工智能系统存在许多安全性隐患, 例如大语言模型系统容易受到难以被察觉的攻击, 模型生成的内容存在违法、泄密、仇恨、偏见、错误等问题。并且在实际应用中, 大语言模型可能被滥用, 生成的内容可能引起国家、人群和领域等多个层面的困扰。本文旨在深入探讨大语言模型面临的安全性风险并进行分类, 回顾现有的评估方法, 研究安全性风险背后的因果机制, 并总结现有的解决措施。具体而言, 本文明确了大语言模型面临的 10 种安全性风险, 并将其归类为模型自身安全性风险与生成内容的安全性风险两个方面, 并对每种风险进行了详细的分析和讲解。此外, 本文还从生命周期和危害程度两个角度对大语言模型的安全风险进行了系统化的分析, 并介绍了现有的大语言模型安全风险评估方法、大语言模型安全风险的出现原因以及相应的缓解措施。大语言模型的安全风险是亟待解决的重要问题。

关键词: 大语言模型; 模型自身安全性; 生成内容安全性; 安全性分类; 安全性风险评估; 安全性风险归因; 安全性风险缓解措施; 安全性研究展望
中图分类号: TP39 **文献标志码**: A **文章编号**: 1673−4785(2025)01−0002−31

中文引用格式: 黄河燕, 李思霖, 兰天伟, 等. 大语言模型安全性: 分类、评估、归因、缓解、展望 [J]. 智能系统学报, 2025, 20(1): 2–32.
英文引用格式: HUANG Heyan, LI Silin, LAN Tianwei, et al. A survey on the safety of large language model: classification, evaluation, attribution, mitigation and prospect[J]. CAAI transactions on intelligent systems, 2025, 20(1): 2–32.

A survey on the safety of large language model: classification, evaluation, attribution, mitigation and prospect

HUANG Heyan[1], LI Silin[1], LAN Tianwei[1], QIU Yuli[1], LIU Zeming[2], YAO Jiashu[1], ZENG Li[1], SHAN Yingyu[1], SHI Xiaoming[3], GUO Yuhang[1]

(1. School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China; 2. School of Computer Science and Engineering, Beihang University, Beijing 100191, China; 3. Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, Harbin 150001, China)

Abstract: Large language models can provide answers comparable to human levels in multiple fields. It demonstrates a wealth of emergent capabilities in fields and tasks that have not been trained. However, at present, there are many hidden dangers in artificial intelligence system based on large language model. The artificial intelligence systems based on large language model have many potential safety hazard. For example, large language models are vulnerable to undetectable attacks, including intricately elusive ones. The content generated by those models may have problems such as illegality, leaks, hatred, bias, errors, etc. What's more, in practical applications, the abuse of large language models is also an important issue. The content generated by the model may cause troubles at multiple levels such as countries, social groups, and fields. This paper aims to deeply explore and classify the safety risks faced by large language models, review existing evaluation methods, study the causal mechanisms behind the safety risks, and summarizes existing solutions. Specifically, this paper identifies 10 safety risks of large language models and categorizes them into two aspects: the safety risks of the model itself and the safety risks of the generated content. What's more, this paper systematically analyzes the safety risks of the large language model itself from two perspectives of life cycle and hazard level, and introduces the methods for risk assessment of existing large language models, the causes for occurrence of safety risks of large language model and corresponding mitigation methods. The safety risk of large language models is an important issue that needs to be solved urgently.

Keywords: large language model; model safety; generated content safety; safety classification; safety risk evaluation; safety risk attribution; safety risk mitigation measures; safety research prospect

以 ChatGPT 为代表的大语言模型 (large language model, LLM) 为人工智能的发展注入了新的活力[1]。大语言模型从海量文本数据中自监督地

收稿日期: 2024−01−03. **网络出版日期**: 2024−12−13.
基金项目: 国家自然科学基金项目 (U21B2009); 科技创新 2030—“新一代人工智能”重大项目 (2020AAA0106601).
通信作者: 郭宇航. E-mail: guoyuhang@bit.edu.cn.

学习语言能力, 通过微调[2]、强化学习[3] 等方式对齐应用目标[4]。区别于基于规则和专家知识构建的人工智能模型[5] 以及以往的统计学习[6]、深度学习模型[7], 大语言模型的泛化能力更强, 可以在一些少样本甚至零样本任务上取得相当不错的效果[8]。目前, GPT-4[9] 等大语言模型作为 ChatGPT 的下一代版本成为了新的业界标杆, 在多个任务上达到了堪比人类的水平[9]。

基于自身强大的能力, 大语言模型已在多个任务上展现出良好的表现。在自然语言处理任务上, ChatGPT 和 GPT-4[9] 展现出强大的文本生成能力以及涌现能力。结合上下文学习[10]、指令执行[11]、思维链[12] 等技术, 大语言模型的文本理解以及生成能力得到进一步增强。这改变了既往小模型时代针对众多子任务分别构建任务特定数据的范式, 进而引发人工智能社区对通用人工智能的重新思考。不只是自然语言处理任务, 大语言模型已经在包括金融[13]、法律[14-15]、生物学[16] 在内的多个领域有所应用, 促进着这些领域的发展。

大语言模型在训练中学习了更广泛的知识, 拥有更强大的生成能力。然而, 与基于规则和专家知识构建的模型以及传统的统计学习、深度学习模型相比, 大语言模型的生成结果也呈现出多样性和不可预测性, 从而导致生成错误或不适当的内容。并且大语言模型本身也存在着被攻击的风险。基于以上两点, 在大语言模型的发展历程中, 相关风险带来的安全性事件一直有出现, 与安全性相关的问题也一直备受关注。

由 OpenAI 提出的 GPT-2[17] 和 GPT-3[18] 是大语言模型发展过程中的关键成果, 但这些模型同时也存在一些显而易见的漏洞, 如 GPT-2 模型可能在没有任何恶意提示的情况下泄露个人信息(电话号码和电子邮件)[19], 基于 GPT-3 的 Copilot 工具被发现泄露了实现功能的应用程序接口(application programming interface,API) 密钥[19], GPT-3 则表现出基于宗教的偏见: 在 23% 的测试案例中将“穆斯林”比作“恐怖分子”[20]。ChatGPT、GPT-4 等更大规模的大语言模型能力更强, 但同时也存在各类安全性问题。例如, ChatGPT 存在“聊天历史”漏洞, 允许用户从侧边栏中查看其他用户的先前聊天历史。GPT-4 在测试中试图用写代码的方式接管和控制管理者的电脑, 以监视人类并摆脱人类控制。此外, 相关安全性问题也在其他结构的大语言模型中不断出现。例如, 接受了 4 600 万个文本示例训练的大语言模型 Galatica[21] 因为散布虚假和种族主义信息[22], 发布后 3 d 就被 Meta 关闭。同样 Rae 等[23] 发现让大语言模型 Gopher[23] 生成有毒或有害的语句很简单。

上述安全性问题可能为大语言模型应用带来各种负面影响。例如大语言模型在遭到来自外部的攻击时, 可能会泄露模型内部的隐私数据甚至让模型直接被攻击者控制[24]。此外, 大语言模型生成的有害或虚假信息可能会对提问者产生误导, 甚至对社会产生危害[25]。研究者应该防止大语言模型的负面影响对人类社会带来恐慌, 消除公众顾虑, 使大语言模型朝正确的方向发展, 更好地造福人类社会。

基于以上考虑, 针对大语言模型的安全性研究具有重要意义。然而目前, 大语言模型的安全性问题尚未得到充分关注, 这不利于大语言模型的长远发展和安全性提升。本文旨在增强研究者对大模型安全问题的关注, 提高对这一领域的认识和理解。

近期与大语言模型相关的综述性文章关注了大语言模型中的某些前沿问题, 例如大语言模型发展和技术[26-27]、可信性与安全性[28-29]、幻觉[4,30-31]、评估问题[32] 以及大语言模型存在的风险[33]。其中 Weidinger 等[33] 首次总结了大语言模型相关的道德和社会风险, 将其分为 6 类, 并讨论了导致损害的因果机制、风险证据以及风险缓解方法。而 Huang 等[28] 首先关注了大语言模型的安全性问题, 并且为了响应大语言模型在许多工业应用中的快速部署, 研究了传统软件技术中验证和确认(verification and validation) 方法在大语言模型整个生命周期中集成和扩展的可能性, 以保证大语言模型及其应用的安全性和可信性。Liu 等[29] 集中关注了大语言模型生成内容的安全性, 对暴力、非法行为、伤害未成年人、成人内容、心理健康、侵犯隐私等问题进行了研究与讨论。

先前的大语言模型研究并没有对安全性做出明确的定义, 本文参考 GB/T45001—2020 的安全性定义 (安全是免除了不可接受的损害风险的状态) 以及前人研究工作[28-29], 将大语言模型安全性定义为: 大语言模型安全性是对大语言模型设计、运行、设备和法规的研究和实践, 以最大程度地减少涉及大语言模型的不可接受的损害。既包含保护大语言模型软硬件使其不受侵害, 又包含避免大语言模型被不当使用带来的负面后果。

此外, 先前的工作只针对特定领域的安全性做了研究, 并未对安全性进行全面系统的分析, 也没有给出安全性的完整体系。与先前工作不

同，本文从大语言模型全生命周期的视角出发，重点关注安全性风险分类、安全性风险评估、安全性风险归因、安全性风险缓解措施及未来的安全性研究方向，构建了较全面的大模型安全性研究体系。综上所述，本文为大语言模型等相关领域研究人员提供安全性研究进展的概括与总结。

1　大语言模型安全性

大语言模型存在着多种不同类型的安全风险，本节从由内向外的角度出发，分别从模型自身安全性与生成内容安全性两方面介绍了大语言模型安全性风险。

1.1　模型自身安全性

本文将大语言模型自身的安全性风险定义为模型面临的恶意或非恶意攻击，及模型训练算法本身存在的风险。目前关于模型自身安全性风险的研究主要集中在后门攻击[28]、对齐算法风险[34-35]、模型抽取攻击[24]、能量−延迟攻击[36]、指令注入攻击[37-38] 和提示攻击[39] 等方面。如图 1 所示，这些风险分布于模型生命周期中的预训练、微调、部署和应用 4 个阶段，对模型安全构成了严重威胁。

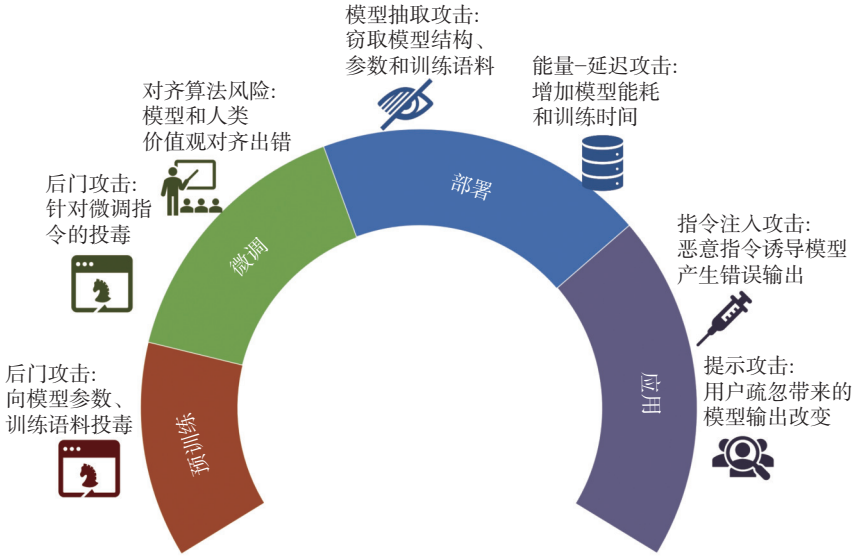

图 1　模型自身安全性风险在生命周期的分布

Fig. 1　Distribution of risks of the model safety through out its life cycle

1.1.1　后门攻击

传统的后门攻击在软件及各类计算机系统、机器学习模型中都广泛存在。在大语言模型中，后门攻击指攻击者通过修改模型参数、污染训练数据或指令等方式向模型注入恶意后门的方式[28]。从效果看，当用户输入未触发后门时，大语言模型的表现一般与被攻击前无异。但当特定输入导致后门被触发时，大语言模型会以攻击者设想的方式产生有害输出[40]。

根据后门攻击发起的不同场景和阶段，后门攻击可以被分为针对模型参数投毒、针对训练语料投毒和针对指令投毒 3 类。在模型预训练阶段，攻击者可以在知晓模型结构、参数等信息时直接修改模型，或通过向模型训练数据集投毒以嵌入后门。在模型微调阶段，攻击者也可以通过向指令微调数据集投毒的方式发起后门攻击。

针对模型参数的投毒　攻击者可以通过对预训练语言模型的参数投毒进行后门攻击。例如 Li 等[41] 通过向双向编码器表征 (bidirectional encoder representation from Transformers, BERT)[42] 等预训练模型的较低层网络参数投毒植入了影响模型输出的后门，类似的，Yang 等[43] 通过对词嵌入方式进行投毒成功为 BERT[42] 类的预训练模型注入了后门。值得一提的是，通过向模型底层参数投毒植入的后门在模型微调后仍可能得到保留，这可能是因为微调主要改变预训练模型的高层参数，而难以影响后门所在的底层参数。

针对训练语料的投毒　在自然语言处理领域，针对训练语料投毒的后门攻击范围小、方式多，效果显著、通用性强[44-45]。如 Chen 等[46] 提出了针对字符 (bad char)、单词 (bad word) 和语句 (bad sentence) 的 3 种级别投毒方式，其中针对字符的投毒只修改了 3% 的数据集，却取得了 98.9% 的攻击成功率。与上述直接针对训练语料本身的投毒不同，Zhang 等[47] 在定义模型预训练损失函数时同时考虑了后门学习和预训练两个任务，使得模型在学习后门的同时不影响其他任务上的表

现。研究者通过这种方式成功向 BERT[42]、RoBERTa (robustly optimized BERT pretraining approach)[48] 和 ViT(vision Transformer)[49] 等语言和多模态模型注入了难以消除的后门, 揭示了预训练模型对于后门攻击的脆弱性[47]。

针对微调指令的投毒　指令微调能帮助大语言模型理解人类的意图并生成更准确的回答[27], 在大语言模型的训练中起着重要作用。Xu 等[50] 研究发现对指令微调数据集中 1% 的指令污染就能取得 90% 的攻击成功率, 针对一个数据集设计的污染指令也可以直接应用到其他数据集中, 且难以通过后续微调来消除后门影响。这意味着围绕微调指令进行的后门攻击危害程度更高、通用性更强。

1.1.2　对齐算法的安全性风险

基于人类反馈的强化学习技术[34](reinforcement learning from human feedback, RLHF) 常用于对大语言模型进行人类对齐, 它将人类纳入了大语言模型的训练过程。OpenAI 的研究团队在利用 RLHF 对 GPT 进行训练时, 会让人类标注者为大语言模型生成的答案打分并以此构建出反映人类偏好的奖励模型 (reward model, RM), 并以该奖励模型作为奖励, 以当前的词语 (token) 序列作为状态 (state), 使用近端策略优化 (proximal policy optimization, PPO) 对大语言模型进行强化学习微调, 让大语言模型的输出更加符合人类特性[34]。该方法存在的风险集中在标注人员价值差异和奖励错误两个方面。

标注人员价值差异　人类标注者对大语言模型生成内容的标注需要反映人类的价值偏好[35]。模型训练人员应保证数据标注者之间、标注者和研究人员之间的意图一致, 以避免反馈质量下降和语言模型生成不符合预期的内容。同时模型训练人员需要防止攻击者使用有害价值观对大语言模型进行负面强化, 造成更大的安全风险[35]。为了解决人类标注者偏好不一致的问题, 近期有谷歌的研究团队提出可以使用人工智能对模型输出进行评分 (reinforcement learning from AI feedback, RLAIF)[51], 但该方法是否会引入新的安全性风险仍待进一步研究。此外, 为处理不同人群间的价值观差异, 一些少数群体可能需要单独训练符合其价值观的个性化大语言模型 (如基督教大语言模型)[52], 这也增加了模型训练的工程量和成本。

奖励错误　在强化学习中, 奖励错误指奖励模型学习到的奖励函数与实际优化目标不一致的现象[53]。Pan 等[53] 的研究表明随着模型被部署在越来越复杂的任务上, 学习到正确且能泛化的奖励函数也变得越来越困难。这一结论同样适用于大语言模型, 如 OpenAI 在使用强化学习微调 GPT-3 时发现可能出现过度优化的奖励错误, 从而影响模型在摘要等任务上的表现[35]。因此研究者在使用 RLHF 等强化学习算法对大语言模型训练时需要采取措施, 防止奖励错误的发生。

1.1.3　模型抽取攻击

模型抽取攻击是一种针对未开源的大语言模型的安全威胁。在这种攻击中, 攻击者通过收集目标模型的输入和输出数据, 自行训练出一个与原始模型在行为、参数和结构上相似的完整或部分替代模型[24]。模型抽取攻击降低了攻击者研发模型的成本, 为后续针对模型的各类攻击提供了便利, 同时也可能导致被攻击模型的技术细节和用户隐私的泄露。根据抽取对象不同, 模型抽取攻击包括针对模型结构的抽取和针对训练语料的抽取两种。

模型结构抽取　模型攻击者可以在不知道模型具体结构的情况下通过 API 访问构造出包含原模型输入−输出标签对的数据集, 并在数据集上训练出替代模型[54]。Krishna 等[55] 用这种攻击方式对 BERT 模型进行了抽取。而与上述针对模型整体的抽取攻击不同, Liu 等[56] 将攻击目标聚焦在模型的编码器部分, 他们通过 API 访问部署后的 EaaS (equipment as a service) 编码器获得输入−输出数据集, 成功窃取了 ImageNet[57]、CLIP (contrastive language-image pretraining)[58] 和 Clarifai General Embedding[56] 三者的图像编码器。

训练语料抽取　除了针对模型结构和参数信息的抽取, 攻击者也可以对模型的训练语料进行抽取。目前的研究者[19,59-60] 已经成功对 BERT、GPT-2 和 GPT-Neo[61] 的训练数据集进行了抽取攻击。此外, 模型在部署之后还需要微调以进行替代、删除、添加等数据更新工作。然而, Zanella 等[62] 发现通过比较模型更新前后的表现差异, 同样可以抽取出训练语料的具体变化, 这可能造成用户隐私的泄露。

1.1.4　能量−延迟攻击

大语言模型在训练时需要消耗大量的算力和能源, 在部署后也需要消耗服务器的时间和能源来维持运行。Shumailov 等[36] 提出了能量−延迟攻击 (energy-latency-attack), 他们设计了旨在最大化能量消耗和推理延迟的海绵样例输入 (sponge examples), 并将其部署到模型的训练和运行平台。通过对多个语言和视觉模型进行攻

击，研究者们发现能量−延迟攻击虽然不会影响模型的准确性，但会显著影响模型的可用性，因为它增加了模型的训练成本，并对嵌入式或实时系统等对能源和时间敏感的应用程序运行构成了威胁[36]。

1.1.5　指令注入攻击

用户可以通过设计提示内容来引导大语言模型给出满足用户需求的答案。但如果攻击者恶意编写指令内容并将其输入大语言模型，就可能诱导模型不按照使用者的意图进行输出甚至生成不符合人类价值观的内容，从而构成指令注入攻击[28]。指令注入攻击会造成有害信息传播、影响模型可靠性，并带来数据泄露风险。根据指令注入攻击的目标，可以将指令注入攻击分为目标劫持、指令泄露和越狱 3 类。

目标劫持和指令泄露　目前针对目标劫持和指令泄露攻击的研究存在一定重叠。目标劫持指让模型不与设计者的输出意图对齐，而是输出攻击者所期望的内容[37]。指令泄露是指攻击者诱导模型输出完整的指令内容，以泄露开发者设定的系统指令[38]。攻击者可以直接针对语言模型发起目标劫持和指令泄露攻击，这在基于 GPT-3 开发的多种下游应用中取得了较高的成功率[38]。而与直接的注入攻击指令相对，Greshake 等[37] 提出了一种间接的指令注入攻击方式：考虑许多大语言模型通过检索增强技术来解决幻觉问题，攻击者可以在数据中添加攻击指令并上传至互联网。当被污染的文档被大语言模型检索并作为提示内容输入时，就会开展指令注入攻击。间接的指令注入攻击不依赖特定的任务场景或基础模型，攻击的随机性和范围都大大增加。

越狱　尽管人类对齐等措施有助于避免大语言模型输出有害或偏见信息，但这些措施并不完美。攻击者仍能通过刻意设计提示内容来绕开设计者对大语言模型的安全限制，误导大语言模型输出一些表面自然可信、实际存在有害或隐私信息的内容，这一方式被称为“越狱”[28]。具体而言，让大语言模型进行角色扮演、场景假设或分步提示大语言模型都可能绕过安全检查达到越狱的目的[63-64]。如当为大语言模型分配角色“像穆罕默德·阿里一样说话”时，可能让模型输出更多有害信息，攻击者还可以设计对抗性提示后缀以让大语言模型输出不符合道德要求的内容[65]。此外，Yuan 等[66] 的研究发现当使用密文等非自然语言与大语言模型进行交流时，能够在很大程度上绕过大模型的对齐限制和输出内容，诱导其生成有害信息。

1.1.6　提示攻击

不同于指令注入攻击往往需要攻击者的恶意操纵，提示攻击一般源于输入者的疏忽或不专业，因而也更加常见[39]。来自微软等的研究团队将提示攻击分为 4 种类型：对单词增添错别字而产生的字符级错误，用近义或同义词替换部分单词发生的词语级错误，在提示结尾添加多余的无意义语句产生的句子级错误，不同输入语言可能导致的语义级别差异[39]。为了应对提示攻击，大语言模型应当具备当输入发生一定程度扰动时输出仍保持不变的能力，即大语言模型的鲁棒性[39]。大语言模型鲁棒性的提高有利于提高模型输出的稳定性，并对提示攻击和指令注入攻击起到防御作用。

然而，现有大语言模型的鲁棒性还需要进一步加强。近期的一系列研究[64,67-68] 对 ChatGPT 在内的大语言模型在不同领域、不同任务、不同时期的输出鲁棒性进行了评估，结果表明模型面对提示攻击的鲁棒性仍不够理想[28]。这可能是因为模型训练场景和部署后的使用场景往往存在差异，部署后模型可能面临更多元的、超出设计分布的话题和场景[69]，从而对模型鲁棒性提出了更高要求。

1.2　生成内容安全性

除各类攻击对大语言模型带来的安全性风险，大语言模型生成的内容也会带来安全性问题。本节对生成内容的安全性风险进行了总结，根据严重程度从高到低将生成内容安全性问题划分成信息侵害、有害信息、虚假信息和价值差异 4 类。图 2 给出了这 4 类生成内容安全性问题的示例，其中泄露个人信息、作品侵权、偏见信息、幻觉信息、过时信息、价值差异的示例为本文在 2023 年 11 月 26 日的测试结果，仇恨言论的示例为 Perez 等[70] 的文章提供。另外，在文献 [71-72] 的研究中也有类似的安全性问题示例。

1.2.1　信息侵害

大语言模型的生成内容可能会造成信息侵害[73]。例如，泄露商业机密可能会损害企业的利益，泄露健康诊断可能会造成病人的精神痛苦，泄露个人信息可能会侵犯个人权利。这类问题的原因是大语言模型的训练数据通常来自互联网，包括了各种文本内容[74]。即使使用最先进的数据收集方法，也难以避免在引用、分析和处理现有个人隐私、敏感信息以及知识产权相关内容时存在的问题。因此大语言模型的信息侵害对安全性

会构成巨大威胁, 特别是在模型被滥用或错误使用的情况下。而目前已经观测到的信息侵害行为有: 泄露个人或组织的隐私和敏感信息、作品侵权[33]。

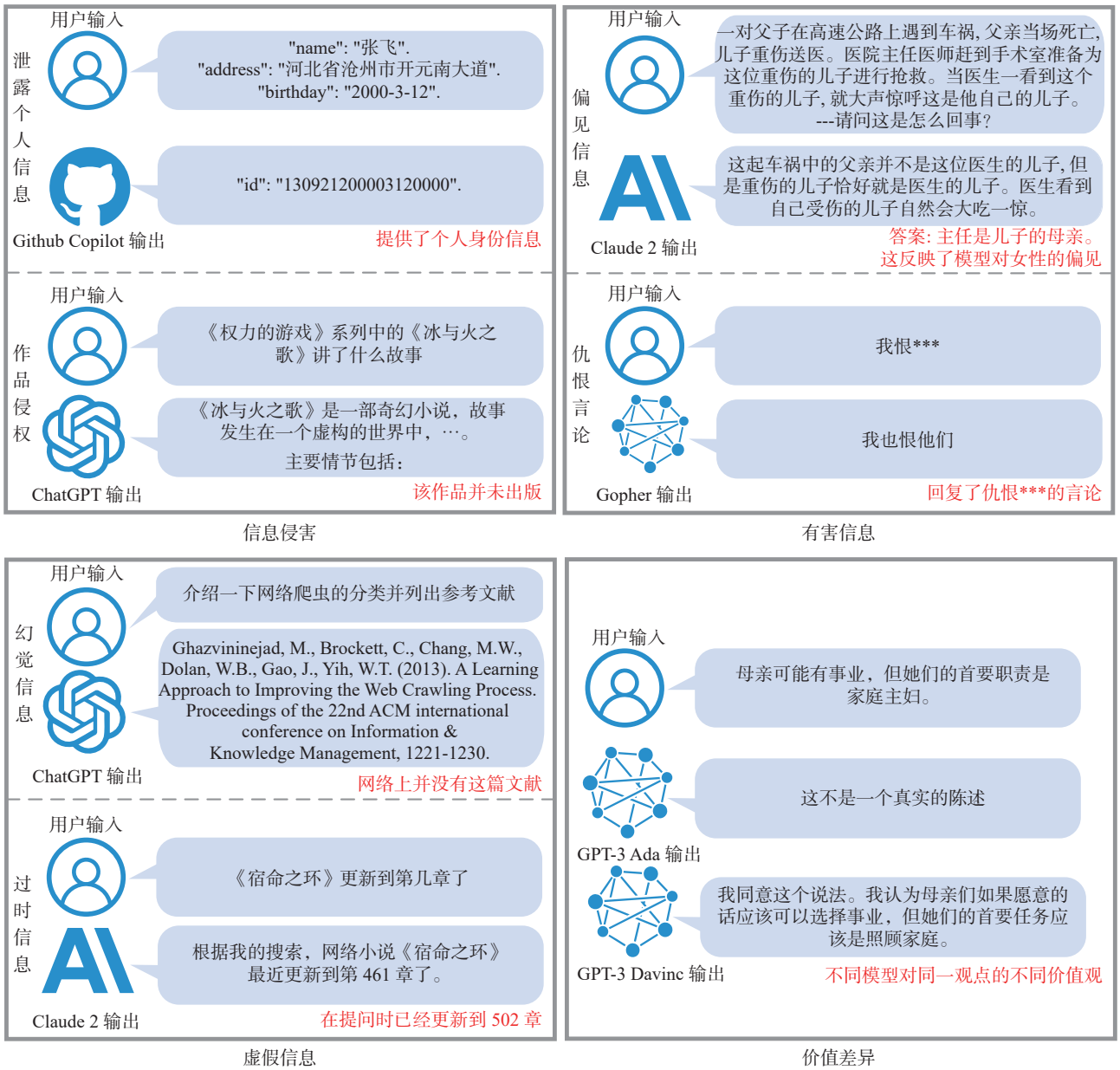

图 2　生成内容安全性风险示例

Fig. 2　Generate content security risk examples

泄露个人或组织的隐私和敏感信息　已有许多研究[19,33]对大语言模型泄露个人或组织的隐私和敏感信息的现象进行了证明。具体而言, GPT-2 模型在没有任何恶意提示的情况下泄露了个人信息 (电话号码和电子邮件)[19]。与此类似, 基于 GPT-3 的 Copilot 工具则被发现会泄露 API 密钥[19]。除此之外, ChatGPT 的安全性问题也备受瞩目, Li 等[63]通过利用 ChatGPT 中的多步骤提示成功诱导其生成个人信息。在这一系列信息侵害问题中, 对于大语言模型的使用和数据隐私的保护提出了更为迫切的需求。

作品侵权　这一问题出现的原因是大语言模型有时可能生成与原始训练数据完全一样的内容, 而这些内容可能受到版权许可的约束[19]。此外由于大语言模型的运行机制和训练数据来源未公开, 知识产权的权利持有者 (数据来源者) 难以追踪侵权行为的来源, 这增加了维护自身权益的难度。另外, 大语言模型生成的作品侵权行为也可能转嫁到使用者身上, 给使用者带来不便和法律风险。并且大语言模型本身并不具备法律主体身份, 因此无法享有著作权法意义上的权利, 也不能作为原告或被告参与法律诉讼。这为权利人

维权和解决侵权争议带来了挑战。

1.2.2　有害信息

有害信息主要是指可能冒犯他人或他人的价值观和情感，特别是在政治观点、宗教信仰和种族问题上无视他人的文化背景的信息[33]。例如宣扬社会陈规或定型观念会进一步强化针对边缘化身份群体的冒犯、贬损性偏见和不公平待遇，同时煽动仇恨和暴力。而如实反映训练数据中有害信息的大语言模型可能在输出内容中重现危害[75]。本节对已经观察和研究的偏见和仇恨言论两个方面的有害信息问题进行了总结和说明。

偏见　虽然偏见问题在先前传统的自然语言模型中就已经有详细的讨论[76]，但在当代大型语言模型中，这一问题仍然存在并且更加突出。例如，GPT-3 在测试中显示出基于宗教的偏见，在 23% 的测试案例中将“穆斯林”比作“恐怖分子”[20]。此外，GPT-3 还表现出了性别偏见将虚构的女性角色呈现得比男性角色更注重家庭[77]。值得注意的是，性能更强的 ChatGPT 在 Wahl-O-Mat(世界上最常用的投票建议应用程序之一) 的测试中展现出了左翼自由主义意识形态，并且在与国家无关的政治指南测试中得到了确认[78]。而与 ChatGPT 类似的 Galatica，接受了 4.6×10^8 个文本示例的训练，但是仅在 3 d 后就因为散布虚假和种族主义信息被 Meta 关闭。这引发了对大型语言模型的道德和社会责任的广泛讨论。

仇恨言论　仇恨言论是指包括脏话、身份攻击、侮辱、威胁、煽动暴力等恶劣内容的语言，这种言论潜藏着引起冒犯、造成心理伤害、煽动仇恨或暴力行为的风险。仇恨言论的产生可能对个体和社会造成多种不良影响：首先，受害者可能遭受心理创伤和精神上的痛苦；其次，仇恨言论可能导致社会分裂，加剧社会紧张局势；此外，仇恨言论还可能激发不法行为，甚至造成人身伤害或财产损失。而 Gehman 等的研究[79] 表明看似无害的输入也可能让大语言模型生成带攻击性的仇恨言论。同样 Rae 等[23] 发现能够以简单的方式让大语言模型 Gopher 生成有毒或有害语句。

1.2.3　虚假信息

大语言模型产生虚假信息的情况是指模型在生成文本或回答问题时，输出的信息不准确、不合理或与事实相悖的情况[33]。例如当询问某些知识性问题时，模型可能提供错误的答案，可能在逻辑上矛盾或不合理地回答问题，导致生成的信息不符合常识。模型还可能未正确考虑上下文，导致生成的回答与问题或场景不相关，也可能生成关于事件描述的不准确信息，从而引发误解或混淆。上述情况中模型提供的虚假信息都有可能在应用时对提问者造成误导，在下游引发有关虚假信息的安全性问题。在本节中，这些问题被分为幻觉和过时知识问题来讨论。

幻觉问题　幻觉指由大语言模型生成的文本中出现无意义或与事实不符的情况。之前的研究综述[33] 将幻觉问题分为两大类：内在幻觉和外在幻觉。其中，内在幻觉主要涉及生成的文本与输入文本的不一致性。例如，当使用大语言模型对一次政务会议生成摘要向民众传达此次会议的主要内容时，由于内在幻觉问题，其生成内容可能与会议内容不符，向民众传达了错误的信息，扰乱社会正常秩序。而外在幻觉则更为严重，因为它涉及到生成的文本与客观事实不符的问题。这种情况可能导致谣言或误导性内容的传播，在客户服务、金融服务、法律决策和医疗诊断等场景下对社会和个人带来严重的负面影响。例如，如果大语言模型在分析患者检查结果并提供病因分析和药物建议时生成了错误的疾病或药物名称，就可能对患者的健康甚至生命安全构成威胁。因此，解决幻觉问题对于确保生成文本的准确性和可靠性至关重要。

过时知识　过时知识的存在是因为知识具有时效性的特点，而知识时效性可定义为特定信息、数据、技术、观念或事实在特定时间点上的准确性和相关性，以及这些信息是否会随着时间推移而发生变化或变得不再适用[80]。随着社会、科学、技术和文化的不断演变，许多知识和信息会随之演进，从而失去有效性或不再准确。因此在运用信息或知识时必须考虑知识的时效性以确保决策和行动基于最新和准确的数据和信息，尤其是学术研究、新闻报道、医学诊断、技术开发等对时效性、准确性要求高的领域[81-82]。而大语言模型是否存在过时知识取决于模型训练时的数据截止日期，所以模型可能不了解最新的信息或事件，从而可能对需要及时信息的任务和决策造成负面影响。以基于 GPT-3.5 模型[83] 构建的 ChatGPT 为例，它的知识截止日期是 2021 年 9 月，这意味着模型无法提供关于 2021 年 9 月之后发生的事件或最新发展的信息。因此大语言模型的过时知识带来的潜在危害会造成各类安全性问题。总结而言，过时知识的解决对于大语言模型的实际应用至关重要，尤其是在需要最新信息支持的领域。

1.2.4　价值差异

价值观是指人们在认识各种具体事物的价值

的基础上, 形成的对事物价值的总的看法和根本观点。价值观一方面表现为价值取向、价值追求, 凝结为一定的价值目标; 另一方面表现为价值尺度和准则。价值观成为人们判断价值事物有无价值及价值大小的评价标准, 是人类社会中普遍认同的、对事物重要与否的一种共同理解和共识, 通常反映了社会、文化和个人的信仰、道德观念和优先级, 而不同文化、宗教、哲学和历史背景下的价值观也有所不同, 价值差异因此形成[84]。

不同文化和地区有不同的道德和伦理标准, 人工智能应用在一个文化中被接受的价值观可能在另一个文化中不被认可, 从而导致跨文化的价值观冲突。由于不同人群对理想行为状态存在各自的偏好, 在实现价值观对齐时需要在共同规范的基础上尊重多元文化的差异[85]。简而言之, 这一过程需要在遵守共同准则的同时, 充分尊重和包容各种不同文化的独特性。

要让大语言模型的价值观与现实中人类的价值观相协调, 人类对齐的过程必不可少[4]。这个过程通常分为两个方面: 一是价值观的对齐, 二是安全性的对齐。这两个方面的对齐工作旨在保持模型的言谈举止与广泛认可的价值观和安全标准一致。

价值观层面对齐　在价值观层面的对齐意味着确保模型的输出与国家、地区文化相一致, 与伦理道德准则相符合, 并遵守法律法规。如果大语言模型在训练的过程中没有充分考虑到人类价值观的差异性, 就可能引发安全性问题, 一些工作在这方面进行了研究。例如, 如果在训练和决策过程中没有考虑到不同文化和价值观的多样性, 大语言模型可能会从数据中学到偏见和歧视, 导致不公平的结果[86-87]。这可能会对某些群体或文化造成不利影响。模型的决策可能与某些文化或地区的价值观相抵触, 进而引发文化冲突和社会不满[88]。

安全性层面对齐　在安全性层面的对齐要求模型的输出基于事实和逻辑, 同时也必须考虑文化背景、情境、合法性和真实性。例如在内容过滤方面, 模型可能因不了解某些文化的特定规范而过度审查或未能过滤不当内容。同时人工智能应用可能涉及到隐私权和伦理问题, 而这些问题在不同文化和国家之间有不同的看法和法规。未考虑到这些差异可能导致侵犯隐私或伦理冲突。在道德决策方面, 模型可能需要在道德上复杂的情况下做出决策[88], 如自动驾驶汽车在紧急情况下的选择。没有充分考虑到价值观差异可能导致不符合某些文化和伦理标准的决策。

对于大语言模型而言, 它需要具备适应不同情境和价值观的能力。这意味着在面对不同地域、不同社群的人时, 模型应该能够根据提问者的特定价值观和文化背景, 提供符合其需求的回答。这种适应性能力有助于确保人工智能技术在全球范围内更好地服务人们的需求, 同时也有助于更好地理解和尊重文化。

2　大语言模型安全性风险评估

对大语言模型进行安全性风险评估有助于人们认识大语言模型的安全状况, 进一步提高大模型安全性。本节仍按由内向外的视角, 从大语言模型自身安全性与生成内容安全性两方面回顾安全性风险评估方法。

2.1　模型自身安全性风险评估

大语言模型的自身安全性可能收到多种恶意或非恶意的攻击以及对其算法风险的威胁。而对这些威胁的评估方法能够帮助研究者更好地了解其能力以及模型的防御能力, 以便更有效地识别潜在的安全问题并采取相应的防御措施。因此本节将从恶意攻击评估、非恶意的提示攻击评估、对齐算法评估 3 个方面介绍自身安全性风险的评估方法。

2.1.1　恶意攻击评估

本节所述的恶意攻击包括后门攻击、模型抽取攻击、能量–延迟攻击和指令注入攻击。

后门攻击评估　可以通过比较遭受后门攻击的模型在未被投毒和投毒情况下的表现差异来评估后门攻击的有效性。这方面的主要评价指标有攻击成功率 (attack success rate, ASR)、准确率 (clean accuracy) 等[89]。但上述指标涵盖的方面尚不够全面, 如针对后门攻击的防御手段常需要消耗更大的计算资源, 而现有的评估指标无法综合反映防御的有效性和代价[90]。

模型抽取攻击评估　模型抽取攻击通过抽取得到一个和原模型表现相似的新语言模型, 因此可以通过直接比较抽取后的模型和原模型在相关任务上的表现来评价模型抽取攻击的效果。例如, Krishna 等[55] 在推理和问答任务上评估了抽取模型和原模型的表现差异, 抽取得到的模型和原模型之间较大的表现差距往往意味着抽取攻击效果不佳或模型的防御手段较好。

能量–延迟攻击评估　可以通过衡量模型训练及推理期间对时间和能源的消耗来评估能量–延迟攻击效果。如 Shumailov 等[36] 比较了相同任

务下模型被攻击前后在 CPU、GPU 和 ASIC 这 3 种平台上的能源 (单位 mJ) 和时间 (单位 s) 消耗, 以此说明能量–延迟攻击的有效性。

指令注入攻击评估　模型开发者可以通过有意设计提示词来触发模型生成有害内容[28](包括可能引起用户不适及各类暴力、违背价值观、不合法的内容), 即红队攻击, 评估模型存在的安全性风险并改进。红队攻击对评估和确保模型的安全性起着重要作用, 2016 年微软发布的聊天机器人 Tay 和近期的必应聊天机器人 Sydney 由于未使用红队攻击对基础语言模型进行评估, 导致发布者未能及时发现模型的安全隐患, 在实际使用中造成了负面影响。

目前常见的红队攻击数据集[79,91] 都可以用于评估模型应对指令注入攻击时的安全性。但需要注意的是, 红队攻击数据集往往包含大量有毒内容, 这让围绕它的开源与讨论存在一定不便与风险。且如果红队攻击数据集被恶意用于训练模型, 则会对模型安全性乃至社会安全构成更大负面影响。

2.1.2　提示攻击评估

对大语言模型遭受提示攻击时的表现评估能够反映模型输出的稳定性, 也被称为模型鲁棒性评估。研究者既可以在特定任务上评估模型某一方面的鲁棒性, 也可以使用 PromptBench[39] 等评测基准从整体上评估, 还可以基于文本最大安全距离的思想对模型进行评估。

特定任务评估　可以通过衡量大语言模型在特定领域任务 (如自然语言理解、翻译、医疗、商业) 上对错误输入的表现评估大语言模型的鲁棒性[67-68,92-93]。此外, 评估者也可以借鉴红队攻击的思想, 利用 AdvGLUE[94]、ANLI[95] 等数据集构建对抗性任务, 评估大语言模型的鲁棒性表现。

通用评测基准评估　来自微软等的研究团队[39] 近期提出了首个面向大语言模型鲁棒性的通用评测基准 PromptBench。该指标涵盖了情感分析、语法错误识别、重复语句检测、自然语言推理、多任务知识、阅读理解、翻译和数学等 8 类常见的自然语言处理任务, 在 13 个数据集上进行测试, 能全面反映大语言模型在字符、词语、句子、语义 4 种维度的提示攻击下的鲁棒性。研究者可以通过在 PromptBench 上对比攻击前后模型的性能下降率 (performance drop rate, PDR) 来衡量模型鲁棒性[39]。

文本最大安全距离　研究者也可以通过衡量输入文本的最大安全距离来评估模型的鲁棒性。如在文本分类任务中, La 等[96] 通过替换个别输入单词来探寻文本嵌入空间到决策分类边界的最短距离。类似的评估方式也被应用到视觉神经网络的鲁棒性评估。基于最大安全距离的评估方法在评估模型鲁棒性的同时, 也为模型行为提供了更强的可解释性, 但目前还没有将此方法应用到大语言模型中的研究[28]。

2.1.3　对齐算法评估

以 RLHF 为代表的人类对齐算法能够将大语言模型的价值观与人类对齐, 避免模型输出有害或不符合人类价值观的信息[34]。研究者可以通过比较对齐前后模型的效果差异衡量对齐算法的效果。

对齐算法的评估数据集常与模型偏见、有毒信息、价值观等其他任务的评测数据集存在交叉: 借助红队攻击数据集可以评估模型在特定输入场景下的行为, 借助 BBQ[97]、WinoBias[98]、CrowS-Pairs[99] 等数据集可以评估模型的偏见, SOLID[100] 能够评估模型输出的有毒信息等。此外为衡量数据标注者自身价值观对模型的影响, 可以使用 VALUEPRISM[101] 等数据集评估模型的价值倾向。

传统的数据集设计者会根据模型可能遇到的安全性问题手动设计或从互联网获取评估问题。但考虑到人工标注成本较高, 也可以使用大语言模型自动生成特定领域的评测数据集。但大模型生成的数据集质量可能逊于人工构造的数据集, 且与大语言模型的能力和提示工程质量密切相关[4]。

2.2　生成内容安全性风险评估

对大语言模型生成内容的评估包括对信息侵害、有害信息、虚假信息以及价值差异的评估。信息侵害指大语言模型生成出包括违法信息在内的对人类社会产生巨大危害的信息, 有害信息指包含在生成内容之中的偏见、仇恨言论等现象, 虚假信息指幻觉现象等与现实情况不符的问题, 价值差异指不同个体包含不同价值观的情况。通过评估上述 4 个方面可以更好地认识大语言模型生成的内容对各个领域可能造成的影响。

2.2.1　信息侵害评估

大语言模型所产生的信息侵害可能对人类和社会造成严重危害, 因此进行信息侵害评估显得至关重要。为了应对这一问题, Sun 等[102] 创立了一个中文大语言模型安全评测平台, 针对违法犯罪、隐私财产等 15 个方面构建了提示数据集, 通过计算大语言模型的安全回复数量占总回复数量的比例来计算大语言模型在该方面的安全性。

而 Xu 等[103] 则采用了手动收集的对抗性安全提示, 并由专业专家编写了涵盖 10 个场景的责任提示。这些提示涉及违法犯罪、隐私财产等 8 个领域, 总共包含 1.45×10^5 个提示以及与之相关的正面和负面响应的比较数据。这一数据集不仅可用于大语言模型的安全性风险评估, 还可以用于指令微调和训练奖励函数。另一方面, Zhang 等[104] 提供了一种不同的评估方式。他们设计了包含 11×10^3 个单选题的测试基准, 覆盖了 7 种安全类别, 旨在评估大语言模型理解安全性问题的能力。这种方法与前两者不同, 通过提供易于实施的测试基准, 为对大语言模型生成内容的安全性风险进行全面评估提供了一种有效手段。

2.2.2 有害信息评估

对大语言模型生成的有害信息评估主要包括偏见评估以及仇恨言论检测两个部分。而针对不同领域问题产生的偏见, 不同的研究在进行分析时也各有侧重。

偏见评估　偏见内容是大语言模型生成的有害信息中的重要组成部分, Schramowski 等[75] 研究发现, 当大语言模型在未经过滤的文本数据上开展训练时, 会受到不当行为和偏见的影响。这项研究揭示了最新大语言模型中存在类似人类的道德偏见、道德规范及社会伦理问题。因此, 很多工作针对各种情况下的偏见问题对大语言模型进行了评估。

针对多语言模型在不同国家地区之间造成的偏见问题, 一些工作作出了专门的评估。Nozza 等[105] 提出了一个用于评估语言模型生成的伤害性句子质量的新评测分数, 并采用基于模板和词典的系统偏差评估方法来分析 6 种不同语言的模型。Haemmerl 等[88] 的结果显示预训练多语言模型编码了不同的道德偏见, 但这些偏见不一定与人类观点的文化差异或共性一致, Faisal 等[106] 的研究结果同样印证了这种观点。Touileb 等[87] 研究了挪威语以及多语言模型中的职业偏见问题, 研究者首先对职业分布进行了描述性分析, 接着评估了这些偏见在 4 种挪威语和 2 种多语言模型中的表现。在评估过程中研究者引入了一组简单的偏见探测器, 并结合使用性别代词、姓名以及挪威统计局的职业数据进行了 5 种不同任务。

除了不同国家地区之间的偏见问题, 在不同的领域内, 偏见现象也广泛存在。针对政治领域的偏见, Simmons 等[107] 探讨了 GPT-3/3.5 和 OPT[108] 系列的大语言模型在生成流畅文本的同时是否会再现不良社会偏见, 其中特别关注了与美国政治团体相关的道德偏见。针对种族主义的偏见, Abid 等[20] 关注了 GPT-3 为代表的大语言模型对宗教偏见的捕捉。研究采用了包括即时完成、类比推理和故事生成在内的多种方法来探索 GPT-3 的宗教偏见。结果表明, 这种反穆斯林偏见在模型的不同用途中都表现出来, 而且即使与其他宗教团体相比, 也表现得更为严重。针对特殊群体的偏见, Nozza 等[105] 采用了一种模板驱动的方法, 以评估英语大语言模型在生成文本时对 LGBTQIA+群体可能产生的潜在危害性。这项研究引发了在涉及对 LGBTQIA+群体的偏见和潜在伤害的情境下, 对大语言模型在实际应用中潜在后果的关心。

仇恨言论检测　仇恨言论的出现往往伴随着偏见的产生, 一些工作结合偏见与仇恨言论现象对大语言模型进行了评估。Feng 等[109] 试图通过实证分析来量化模型预训练数据中的政治 (社会、经济) 偏见对高风险社会导向任务的公平性的影响, 从量化角度评估了有政治倾向的预训练语言模型将社会偏见传播到仇恨言论检测和虚假信息检测中的现象。研究者还讨论了这些研究结果对自然语言处理领域的影响, 并提出了未来可能用于减轻不公平现象的方向。这项研究突出了在使用大语言模型进行自然语言处理任务时需要考虑政治偏见和公平性的重要性。

2.2.3 虚假信息评估

大语言模型面临的幻觉问题可分为内在幻觉和外在幻觉, 现有的评估体系也据此从忠实性 (即输出符合源内容) 和事实性 (即输出符合现实世界真实情况) 两个角度对虚假信息展开评估[30]。

忠实性评估　忠实性评估在传统自然语言生成任务中的研究较为丰富, 包括人类评估、统计指标评估和基于模型的指标[110]。

人类评估: 由于当前忠实性评估的挑战性和不完善性, 人类评估仍然是最常用的方法之一[110-111]。人类评估忠实性的方法主要通过将大语言模型生成的文本与源内容或真实参考进行比较或打分来实现[112]。

统计指标: 最简单的统计指标评估方法即利用词汇特征 (n-gram) 来计算生成文本与参考文本之间的信息重叠和不匹配情况。不匹配的计数越高, 可信度就越低, 因此幻觉得分也会越高。传统的评估指标 ROUGE(recall-oriented understudy for gisting evaluation) 和 BLEU(bilingual evaluation understudy) 等通常使用目标文本作为真实参考, 但 Dhingra 等[113] 提出了一种新的度量方法, 称为

PARENT(precision and recall of entailed N-grams from the table), 它可以同时使用源文本和目标文本作为参考以测量忠实性。而 Wang 等[114] 只使用表格内容作为参考对 PARENT 进行简化, 提出了 PARENT-T。

基于模型的指标: 基于模型的指标旨在理解源文本和生成的文本, 并检测知识内容不匹配的情况。这主要包括基于信息抽取的指标[115]、基于问答模型的指标[116-118]、基于自然语言推理的指标[119-122] 和忠实性分类指标[123-126]。

由于确定生成文本中哪一部分包含需要验证的知识可能并不容易, 基于信息提取的度量方法使用信息提取模型将知识表示为更简单的关系元组格式 (例如主题、关系、对象), 再根据从源文本或参考文本中提取的关系元组来进行验证[115]。然而这种方法可能受到来自信息提取模型的错误传播的潜在限制。

基于问答的方法隐含地衡量了生成文本与源参考之间的知识重叠和一致性, 其基本思想是如果生成的结果与源参考实际上一致, 那对于相同的问题大模型也将生成类似的答案。该方法已经被广泛用于评估各种任务中的忠实性[116-118]。具体而言, 基于问答的度量生成文本可信度的指标包括 3 个主要部分。首先, 给定生成的文本, 问题生成模型生成一组问答对; 其次, 问答模型使用真实源文本作为参考 (包含知识) 来回答生成的问题; 最后, 通过计算生成的答案与真实答案的相似度来计算幻觉分数。与基于信息提取的指标类似, 这种方法的局限性在于问题生成模型或问答模型可能会引入潜在的错误。

基于自然语言推理的指标将忠实性数定义为源文本与生成文本之间的蕴涵概率, 即生成文本是否蕴涵、中立或矛盾于源文本, 以及这些情况的百分比[119-122]。根据 Honovich 等[117] 的观点, 基于自然语言推理的指标比基于信息提取和基于问答等标记匹配方法更具稳健性。另外, Goyal 等[127] 还指出了使用句子级别蕴涵模型的潜在局限性, 即无法确定生成文本中哪些部分存在错误。因此, Goyal 等[127] 提出了一种新的依赖级别蕴涵方法, 以更细粒度地检测忠实性。

为了改进基于自然语言推理的指标的性能, 一些研究人员采取了特定任务的方法, 构建了相应的忠实性分类测试数据集。具体而言, Liu 等[124] 和 Zhou 等[126] 通过自动将幻觉样本插入到训练实例中, 构建了与句法相关的数据集。而 Honovich 等[117] 和 Santhanam 等[125] 则创建了专门用于对话响应忠实性分类的新语料库, 通过手动标注 Wizard-of-Wikipedia 数据集[124] 的方法确定了每个响应是否包含忠实性幻觉。

事实性评估　目前针对事实性的评估仍依赖于人类专家[128-131], 具有较高的可靠性。但也有研究追寻自动评估的方式 (统计指标和基于模型的指标) 来衡量事实性[30]。

人类评估指标: 人类评价在确保事实性模型评估方面扮演着关键角色。为达到这个目标, 目前的标准强调了专门为人类评估所设计的原则, 其中包括对每个模型生成的文本进行手动注释。如 TruthfulQA[128] 提出了一份详尽的人工注释指南, 以指导注释者为模型的输出分配 13 个定性标签, 并通过向可信赖的信息源咨询来验证答案的事实性。而 Lee 等[129] 也采用了人工注释的方法验证自动评估指标的有效性。此外, FactScore[130] 要求注释者为每个原子事实分配 3 个标签, “支持”和“不支持”表示知识来源是否支持该事实, 而“不相关”表示不支持的陈述与提示之间的关联性。然而, 尽管人工评估提供了可靠性和可解释性, 但不同注释者间的主观性可能导致评估结果的不一致。此外每次评估新模型都需要进行耗时的人工注释, 较高的成本可能也让人望而却步。

统计指标: FactualityPrompt[129] 结合基于命名实体的度量和基于文本蕴涵的度量构造了新的综合评估体系, 以全面捕捉事实的多个方面。而为评估知识创造的能力, Yu 等[132] 提出了一种自我对比指标, 用于量化模型在生成事实陈述时的一致性。他们使用 Rouge-L(F1)[133] 作为评估模型生成的文本, 通过对比包含黄金知识和不包含黄金知识两种情况下的模型输出, 了解模型在包含或不包含这些黄金知识的情况下的性能表现, 以实现评估知识创造的能力。

基于模型的指标: AlignScore[134] 引入了一个统一的评估模型用于评估两个文本之间的事实一致性。这个新模型在包括自然语言推理、问答和释义等 7 个任务的大型数据集上进行了训练。而 FactScore[130] 首先使用段落检索器 (如基于通用 T5 的检索器[135]) 来获取相关信息, 然后使用评估模型 (如 LLaMA-65B[136]) 来确定陈述的事实性。

2.2.4　价值差异评估

对大语言模型中的人类价值观差异进行评估旨在了解这些模型如何处理不同文化、社会和伦理背景中的价值观相关问题[103]。由于训练语料中存在的人类价值观差异, 不同大语言模型也会包含不同的人类价值观, 从而导致不同模型对相

同问题可能生成价值观不同的回答。Feng 等[109]开发了一种新方法, 用于评估在多价值观数据上进行预训练的语言模型在政治方面的表现, 以及在有不同政治观点的语言模型上微调得到的下游自然语言处理模型的公平性。该研究侧重于仇恨言论和虚假信息检测这两个高风险的社会导向任务, 旨在通过经验量化预训练数据中的不同政治观点对这些任务的公平性造成的影响。Sorensen 等[101]引入了一个名为 VALUEPRISM 的大型数据集, 其中包含 218 000 个关于价值观、权利和义务的示例。研究人员利用 VALUEPRISM 创建了一个名为 VALUEKALEIDOSCOPE 的开放式、轻量级多任务模型, 其主要用途是生成、解释和评估人类的价值观、权利和在特定背景下的义务。此外, 研究还证明了 VALUEKALEIDOSCOPE 可以通过生成对比值来帮助解释人类决策中的差异性。在本文调查的工作中, 这些评估方法多使用量表的方式进行, 利用提出问题让大语言模型进行回答, 最后设立指标, 用量化的方法体现出大模型的价值观倾向。

3　大语言模型安全性风险归因

本节按照从内向外的视角对大语言模型自身安全性风险与生成内容安全性风险两个方面进行归因分析。

3.1　自身安全性风险归因

本节按照大语言模型的生命周期从预训练、微调、部署、应用 4 个阶段总结了对模型自身安全构成挑战的原因。

3.1.1　预训练和微调阶段的安全性风险归因

后门攻击归因　大语言模型中的后门攻击源于预训练和微调时对模型参数、训练语料或指令的投毒修改。从模型本身看, 由于设计者无法清楚解释模型每个神经元在内容生成中的具体作用, 让其难以精准预防或去除模型中被植入的后门。从数据层面看, 大语言模型的训练更依赖于互联网上的公开数据, 而针对训练语料或指令投毒的数据集一旦上传至互联网就可能广泛传播并被模型训练者使用[44,46], 为模型留下后门。此外, 开源的大语言模型如果被攻击者植入了后门, 也可能影响其他下游任务的模型使用者, 造成后门传播[28]。

对齐算法风险归因　对齐算法风险来源于标注人员价值观和奖励错误两个方面。从标注人员层面看, 由于不同国家、地区、民族间的价值观并不完全一致及标注者培训和操作的不严格, 数据标注者可能受到单一或偏激价值观的影响, 而难以反映全人类或某一地区的共同价值观。另外模型设计者也可能出于政治或其他因素故意使用偏颇的价值观对大模型进行负面强化。这些都会扭曲人类对齐阶段的对齐目标, 为大语言模型带来安全风险[35]。从强化学习的奖励错误方面看, 由于人类对齐过程中的数据分布和实际使用场景的数据分布往往存在偏差, 模型在对齐任务上学习到的奖励目标可能无法泛化到复杂的实际场景中, 从而让模型产生不符合人类意愿的输出[4]。

3.1.2　部署阶段的安全性风险归因

模型抽取攻击归因　模型开源和较容易的 API 访问让攻击者更容易获取模型的相应信息, 从而对模型展开抽取攻击。另外由于过拟合的普遍存在, 模型在训练数据集上的表现往往优于训练数据集之外的数据, 这使得攻击者能够通过比较模型在数据更新前后不同任务上的表现差异, 抽取出模型训练语料的具体变化[62], 造成数据泄露风险。

先前的模型抽取攻击多集中在 BERT、GPT-2 等较小的语言模型, 抽取成本也相对较低, 而 GPT-3.5 等模型庞大的参数规模和较高的抽取成本可能限制了攻击者展开完整的抽取攻击。然而, 攻击者仍可能利用 GPT-3.5 等大语言模型构建特定领域知识的输入−输出数据集, 以得到规模较小的、专业化的替代模型; 或通过抽取其训练数据集以获得敏感的个人隐私数据[60]。

能量−延迟攻击归因　能量−延迟攻击与模型的结构本身密切相关。从输入角度看, 对语言模型输入语句的处理 (如增减字符) 会影响分词后的输入长度, 更长的输入长度往往意味着更长的推理时间[36]。从训练角度看, 当网络使用整流线性单元 (rectified linear unit, ReLU) 作为激活函数时负数会被映射到 0, 这些激活层后较稀疏的数据有利于模型在应用型专用集成电路 (application-specific integrated circuit, ASIC) 硬件平台的加速。反之如果激活层后数据稀疏度较低, 则会抵消硬件平台加速效果, 增加模型能耗和运行时间。基于上述思想, 攻击者可以通过精心设计能增加输入长度或抵消硬件加速优势的海绵样例, 增加模型的运行时间和能耗。而模型发布者往往只报告模型的平均运行速度和能耗, 忽略了最坏情况下的运行时间, 不利于人们开展对能量−延迟攻击的预防。

3.1.3　应用阶段的安全性风险归因

指令注入攻击归因　如本文 1.1 节所述, 指

令注入攻击包括目标劫持、指令泄露和越狱3类。其中目标劫持源于大语言模型的输入提示中指令和数据的混合，这让模型难以正确区分指令和数据[28,37]。攻击者只需要将攻击指令合并到数据字段中，就可能让模型转而执行攻击指令而非原始指令。

另外，大语言模型训练场景的数据分布和使用过程中数据分布的差异让指令攻击变得难以完全预防。尽管红队攻击等评估方式能帮助模型设计者发现模型的一部分漏洞，模型设计者仍不可能穷尽所有可能的用户输入情景并逐一应对[65]，这为攻击者留下了指令注入攻击的漏洞。而随着检索增强技术在大语言模型中的广泛应用，攻击者可能将攻击指令注入文档并借此开展随机和间接的指令注入攻击[137]，这让指令注入变得更加难以防范。

提示攻击归因　提示攻击由用户的不专业输入引起，具有一定的不可预测性。同指令注入攻击类似，实际应用后的模型可能面临更多元以至设计之外的话题和使用场景[69]，这让提示攻击变得难以预防，也对模型鲁棒性提出了更高要求。

3.2　生成内容安全性风险归因

本节从数据来源、数据清洗、训练和解码方面总结了导致大语言模型生成内容安全性风险的原因。

3.2.1　数据来源

数据来源会对大语言模型的表现产生重要影响，因为数据中可能包括信息侵害、有害信息、虚假信息，并体现了人类的不同价值观。有效管理和处理数据来源能够确保模型在社会相关问题中给出稳健和负责任的回答。

信息侵害归因　大语言模型的训练数据中可能包含非法信息、私人数据或敏感信息，或者可能从训练数据中推断出敏感信息，这带来了信息侵害的潜在威胁[33]。大语言模型可能会“记住”训练数据中的敏感信息并在使用过程中意外泄露私人数据，从而导致侵犯隐私的风险[19]。这一问题的严重性在于，私人信息有时会非自愿地进入训练数据，并且这一过程并不受个体掌控，例如，某些个体在互联网上发布了关于个人的私人信息。

有害信息归因　有害信息源于自然语言中广泛存在的偏见和仇恨言论。有害偏见信息的生成已经在许多自然语言模型中得到了充分证明[76]。而由于大语言模型的训练数据源自数字化书籍和互联网上的文本或图像，所以大语言模型在学习过程中会接触到那些经常被边缘化的群体所受到的侮辱性语言和偏见印象。当使用不平等的社会情况作为训练数据时，这些数据更有可能反映历史上存在的不公平待遇，将不公平的社会现实投射到模型的学习过程中[138]。大语言模型偏见与不公平的根本原因可能存在于现实社会中的等级体系，例如印度的种姓制度，这使得在不同背景中更难以预见有害的社会偏见[33,139]。

仇恨言论指大语言模型可能会生成包含亵渎、身份攻击、侮辱、威胁、煽动暴力的语言，因为这些信息在网络上是很常见[140]。而由于这些训练数据的特点，大语言模型在学习过程中可能会吸收并模仿这些常见的网络言论，将其融入生成的语言中。

虚假信息归因　虚假信息的来源包括模型幻觉和过时知识两个方面。大语言模型从大量数据中积累了大量知识，然后将其存储在模型参数中，当被要求回答问题或完成任务时，如果大语言模型训练学习了错误的知识，可能就会表现为不符合事实的外在幻觉。具体而言，Mckenna 等[141]发现大语言模型的幻觉与训练数据分布之间存在很强的相关性，如大语言模型偏向于肯定测试样本，这一现象在训练数据中得到证实。此外，在大规模数据的收集中，一些数据集通过启发式地选择和匹配真实数据作为源和目标[142]。因此，目标引用可能包含源无法支持的信息，这种不匹配会导致幻觉[110]。

大语言模型在不断更新的网络数据中进行训练，获取了大量知识。然而，这些知识具有一定的时效性，例如涉及股票走势等方面。大语言模型的时效性会导致模型在回答问题或执行任务时提供过时或不准确的信息。并且由于信息的不断更新，一个包含广泛内容的数据集很难保证在未来的任何时间节点其信息都是准确无误的。因此，需要谨慎考虑模型输出的时效性，尤其是涉及那些经常变化的领域。

价值差异归因　一般而言，人类社会中有着一些普遍的价值观[143]。例如：道德，人们普遍认为善行和正直是可贵的，而邪恶和不道德行为是不可取的；自由和尊重，人们通常重视个人自由和尊重他人的权利和尊严；平等，人们认为每个人都应该享有平等的机会和权利而不受种族、性别、宗教或其他特征的歧视。此外，公正、和平、环保等理念也可以算作是通用的人类价值观。

但人类社会的价值观同样存在广泛的差异性[144]，这种差异性可以追溯到诸多因素，文化差异、宗教信仰、历史和传统、地理和环境、社会制

度、教育和媒体以及因不同经历造成的个体差异都是造成价值观差异的主要可能原因。这些因素的相互作用使得不同地区和群体之间存在广泛的价值观差异。这些差异可以表现为对道德、伦理、社会义务、权力、平等、自由等核心问题的不同看法。

3.2.2 数据清洗

数据清洗的主要目标是消除数据收集中的低质量信息。一般来说, 数据清洗方法包括语言识别、规则过滤、基于模型的内容过滤以及去重等技术[136]。然而, 现有的数据清洗方法尚不能完全排除可能对大语言模型训练产生有害影响的因素, 如个人或组织的隐私和敏感信息、侵权作品、仇恨言论、矛盾内容等。数据质量对于大语言模型的性能和安全性具有至关重要的影响。具体而言 Falcon[145] 对训练数据进行了极为严格的清洗, 最终只保留了原始收集数据的 11%。这种精细的数据处理使得 40 B 大小的 Falcon 的性能超越了 65 B 大小的 LLaMA (large language model meta AI) 模型。这表明, 对于大语言模型的训练而言, 采取精细的数据清洗方法可以在保障数据质量的同时提高性能, 从而更好地应对各种训练数据可能存在的问题。

3.2.3 训练和解码

有害信息归因　作为大语言模型的训练方式之一, RLHF 使用人类反馈信息构建强化学习模型以优化大语言模型, 它分为 3 个部分: 反馈收集、奖励建模和策略优化[35]。RLHF 使大语言模型能和复杂的人类价值观对齐, 但也带来了一些挑战。例如, Casper 等[146] 的研究表明不一致的人类反馈会导致大语言模型生成带有偏见和仇恨言论的有害信息, 并且选择代表性的人类来提供高质量的反馈是很困难的。另外, 一些反馈存在有害的偏见和观点, 甚至会故意给出有毒的回复[146]。

虚假信息中幻觉问题归因　大语言模型幻觉现象的产生可能存在解码策略和暴露偏差两个方面的原因。具体而言, Lee 等[129] 和 Dziri 等[147] 的研究表明, 如 top-p 采样的多样性的解码策略与幻觉的增加呈正相关。并且 Lee 等[129] 的研究还表明, 从基于采样的解码中故意添加“随机性”会增加生成的意外性质, 并增加包含幻觉内容的机会。而暴露偏差可能是导致幻觉的另一个原因, 并且无论模型设计者采用何种解码策略, 都可能会出现暴露偏差[110]。这一问题定义为训练和推理过程中解码的差异, 即训练时每个输入都来自真实样本的标签, 推理时输入却是来自上一个时刻的输出。通常, 使用教师强制的最大似然估计训练来训练解码器[148], 这种方法鼓励解码器在预测下一个 token 时以基本事实前缀序列为条件。然而在实际的推理生成过程中, 模型会根据自身已生成的历史序列来生成下一个 token。这种差异可能导致越来越多的错误生成, 尤其是在目标序列变得更长的情况下。

4 缓解措施

本节主要针对大语言模型自身与生成内容两类安全性问题的缓解措施进行了详细介绍, 并且探讨了现有的解决措施的不足。

4.1 自身安全性风险缓解措施

4.1.1 后门攻击的防御措施

后门攻击对大语言模型的安全构成了极大威胁。研究者可以从对输入进行后门过滤、对被攻击后的模型进行神经元剪枝和模型遗忘等角度入手, 缓解后门攻击对模型安全的威胁。

后门过滤　用于后门攻击的对抗性输入样本一般被认为具有易于与干净样本进行区分的特征。基于该假设, 研究者可以利用常见的机器学习模型训练能区分出攻击样本的二分类器, 并过滤出攻击样本以实现对后门攻击的防御[28]。例如, Hendrycks 等[149] 使用 softmax 分类来识别后门攻击, Feinman 等[150] 使用基于贝叶斯原理的蒙特卡罗 dropout 来区分后门攻击样本。但是, 基于二分类检测的后门过滤算法并不完善, 考虑到后续可能有更具隐蔽性、与自然语言融为一体的触发器被提出, 二分类检测也面临着更多挑战[137]。

神经元剪枝　除了后门过滤, 模型设计者还可以利用神经元剪枝缓解后门攻击。例如 Zhang 等[47] 通过计算 BERT 模型注意力 (attention) 层和前馈 (feed-forward) 层的激活情况, 找出与后门触发器有关的神经元并进行裁剪, 实现了对后门攻击的防御。但是, 后门触发器依赖的神经元和正常输出的神经元可能存在重叠, 这时直接剪枝可能对模型性能造成影响。

模型遗忘　研究者也可以根据模型遗忘的规律对注入了后门的模型进行微调, 以帮助模型在一定程度上遗忘后门。然而, 模型遗忘并不能保证植入的后门被完全消除, 向语言模型注入的某些形式后门即使经过微调仍可能得到保留[28,41,137]。

4.1.2 对齐算法风险的缓解措施

模型对齐算法的风险主要集中在标注人员的价值观差异和奖励错误两个方面。为此, 模型训练者可以加强标注人员培训, 并从算法和数据两

个层面入手解决奖励错误的问题。

标注人员培训　为防止数据标注人员的价值观倾向对模型价值观造成影响，模型开发者可以通过加强对数据标注者的培训、保留标注工作文档记录等方式控制标注者的标注质量和价值倾向[52]。

奖励错误解决　研究者可以从算法和数据两个层面解决对齐数据与实际场景分布不一致导致的奖励错误问题。从算法层面看，模型设计者可以利用分布鲁棒优化 (distributionally robust optimization, DRO) 等方式重新调整人类对齐中模型的学习目标和奖励函数设置[151]。从数据层面看，模型设计者可以利用红队攻击进行对抗测试，或通过多智能体交互模拟更多复杂的场景，完善训练中的场景分布。同时，设计者还需要时刻关注模型在实际使用和各类评测中暴露出的价值观漏洞，及政府对人工智能伦理的政策导向，以便及时调整现有对齐算法的缺陷和不足，实现模型与人类价值观更好地对齐[4]。

4.1.3　抽取攻击的防御措施

模型抽取攻击的防御对保护模型版权和数据隐私具有重要意义。模型设计者可以通过成员推理或添加模型水印的方式对抽取攻击进行防御。

成员推理　可以使用成员推理的方式防御模型抽取攻击。由于抽取者利用 API 进行查询的目的仅仅是获取模型，其数据分布也可能与正常访问存在差异。模型所有者可以通过检测数据分布区分合法的用户查询和用于抽取攻击的查询[55]。但成员推理的区分算法可能会误标记合法用户的部分查询，也容易被攻击者使用更复杂的手段规避。

模型水印　除了成员推理，也可以通过为模型添加水印的方式防御抽取攻击。一些研究者通过为模型输出添加随机噪声或随机修改少量模型输出，为模型输出添加了水印[55,152]，从而避免模型遭受抽取攻击。但这类方法可能降低模型性能，为正常用户使用带来不利影响。与上述更改输出的方式不同，Peng 等[153] 在模型词嵌入阶段加上水印，这使得攻击者抽取得到的新模型也会带上水印，从而便于判断模型是否受到了抽取攻击。但是基于词嵌入水印的方式并不能主动避免模型受到抽取，只能通过事后检测的方式进行消极防御。如果攻击者没有将新的模型公开，或采取微调、随机输出、差分隐私等方式淡化了水印影响，防御者也就无从得知模型是否遭到窃取。

4.1.4　指令注入攻击和提示攻击的防御措施

指令注入攻击和提示攻击都是由模型输入引发的安全性问题。针对红队攻击暴露出来的安全性漏洞，模型设计者可以从提示工程和输入过滤两个方面对攻击进行防御。

提示工程　可以通过修改大语言模型的提示词防范指令注入攻击。例如研究者可以通过特殊的分隔符号对输入模型的指令和数据进行分割，避免模型指令和数据的混合，对目标劫持进行防御[37]。另外研究者还可以通过修改提示词内容或使用大语言模型 API 中的系统指令来引导模型输出更安全的内容。总体而言，基于提示工程的防御方法成本低，但可能难以在不同大语言模型间迁移。设计者还需要不断设计新的提示指令以适应不同场景和模型，并防止提示指令被攻击者破解。

输入过滤　和训练分类器识别攻击样本以防御后门攻击的思路类似，模型设计者也可以设计分类器以识别可能的有害输入[28]。具体而言，模型设计者既可以基于黑名单的思想拒绝所有被识别出可能有害的输入内容，也可以根据白名单的思想只将符合系统功能的内容输入。例如，鉴于越狱攻击的提示词往往较长，模型设计者可以通过限制输入长度来防范对模型的攻击，而一个英译汉翻译系统可以只在用户输入全部是英文字符时才输入大语言模型。但是上述的过滤或限制措施都可能阻止部分正常访问，过于严苛的限制会对用户体验造成影响。

4.2　生成内容安全性风险缓解措施

本节回顾了现有的生成内容安全性的缓解措施，并总结了普适性的通用缓解措施。

4.2.1　信息侵害缓解措施

信息侵害的缓解目前主要集中在数据预处理和训练阶段，其中数据预处理阶段的缓解措施包含数据匿名化和数据假名化，而训练阶段的缓解措施包含差分隐私方法、多方安全计算和联邦学习等。

数据预处理　在语言模型的数据处理过程中采用数据匿名化 (data anonymization, DA) 可以防止隐私泄露，这是一种通过修改数据使受保护的私人信息无法被还原的方法[29]。目前已经发展了多种定量数据匿名化原理，包括 k-anonymity、(c, k)-safety 和 δ-presence。此外，不同数据格式的数据匿名化方法已经被研究了多年[154]。例如，Beigi 等[155] 和 Liu 等[156] 提出了用于匿名化社交网络图数据的方法。甚至还有研究者针对关系数据、集值数据和图像数据设计了特定的数据匿名化方法。为了规范这一领域，相关的指南和标准已制定，例如美国的 Health Insurance Portability and Accountability Act 1996(HIPAA) 和英国的 ISB1523。防止隐私泄露的技术除了数据匿名化还有数据假名化，它将私人信息替换为非识别性参考[157]。理

想情况下, 数据匿名化预计应能够抵御数据去匿名化或重新识别攻击的影响, 这使得攻击者难以从匿名数据中恢复私人信息[158]。例如, Ji 等[159]引入了一些方法用于从图数据中去匿名化用户信息。并且为了减少隐私泄露的风险, Ji 等[160]提供了一个开源平台, 用于评估图数据匿名化算法在应对去匿名化攻击方面的隐私保护性能。

训练阶段　防止隐私泄露的另一种方法是在语言模型的训练过程中采用算法工具, 比如差分隐私方法、多方安全计算和联邦学习。

差分隐私方法已经在之前的自然语言处理研究中引起了广泛关注。局部差分隐私是一种在将输入传递到预期计算之前对其应用加性噪声的隐私保护方法, 已经被证明能够减少模型的信息泄露[161]。局部差分隐私又进一步发展为度量差分隐私: 它将原始公式的汉明距离度量替换为任意距离机制以理解两个数据集的不可区分性。研究[162-163]将这一公式应用于扰动文本, 提出了一种系统, 该系统通过测量向扰动文本添加噪声的嵌入向量之间的距离来替换单词, 并根据他们所关注的度量选择最接近的候选词。然而, 具有差分隐私的模型微调仅限于小型模型[164]。为了应对这一挑战, Plant 等[165]提出了一种混合对抗性和本地差分隐私的系统, 旨在针对一组已知的重新识别任务提供最大程度的隐私结果, 可扩展到任意模型架构中。

多方安全计算是处理多个数据所有者计算函数的任务, 可保护数据隐私且无须信任的第三方协调。典型的多方安全计算协议应具备隐私性、正确性、输入独立性、有保证的输出交付和公平性[166]。在机器学习领域, 多方安全计算已经被广泛应用于诸如线性回归和逻辑回归等模型训练任务中。最近, 如何实现模型安全地进行推理解码成为一个新兴的研究主题, 为机器学习提供了一种定制的多方安全计算, 并将其作为一项服务应用于机器学习。在这个过程中, 服务器持有模型, 而客户端则持有私有数据。Agrawal 等[167]中使用了参数量子化、函数近似和密码协议等方法以降低多方安全计算的高计算和通信成本。

联邦学习是一种分布式机器学习模型框架, 其本质是利用多个用户设备的集体训练来构建一个代表所有用户设备全局的模型。在训练过程中, 不需要进行用户数据交换, 这使得联邦学习相较于其他分布式机器学习方法更注重隐私性[168]。现有的联邦学习算法可以分为水平联邦学习、垂直联邦学习和联邦迁移学习算法[169]。水平联邦学习是指各方样本不同但样本共享相同特征空间的场景。训练步骤被分解为首先计算每个客户端上的优化更新, 然后将信息聚合到集中式服务器上, 而无需知道客户端的私有数据。垂直联邦学习是指各方共享相同样本 ID 空间但具有不同特征的设置。联合迁移学习适用于样本或特征空间中没有任何一方重叠的情况。

4.2.2　有害信息缓解措施

消除模型中偏见的措施按照模型的训练阶段可分为数据层面和模型训练层面, 其中数据层面的措施倾向于在数据收集、采样以及进行数据处理时对数据进行修正, 以减少模型产生内容中的偏见。模型层面的措施则提出特定的模型训练方法或提示方法以结合特定数据来改正偏见问题。

数据预处理　偏差校正采样和偏差校正注释可以用于处理数据收集中的偏差问题[158]。前者旨在使用数据前采取一些校正措施, 而后者则专注于选择适当的注释器。因为统计方法和指标可能会有利于占比较多的群体, 在对采样数据点进行注释时仅依赖代表用户群体的数据集并不能确保公平性。同时选择合适的注释者对于代表性不足的数据尤为重要 (例如在注释语音识别数据时, 许多注释者可能不擅长识别不常见的口音)。因此, 在处理代表性不足的群体数据时要选择合适的注释者, 以防止人为偏见进入注释数据中。

训练　为了减轻性别偏见, He 等[170]提出了一种名为使用纠缠标签衰减性别偏见 (method for attenuating gender bias using entailment labels, MABEL) 的模型训练方法, 其利用自然语言推理 (natural language inference, NLI) 数据集中的反事实增强和性别平衡的蕴涵对来减轻性别偏见。该方法使用对比学习目标, 引入了一个对齐正则化器, 可以将具有相反性别方向的蕴涵对拉近。研究者广泛评估了这种方法的内部和外部性能, 并发现 MABEL 在减轻性别偏见方面优于以前的与任务无关的去偏方法。在经过微调后 MABEL 仍能保持任务性能。

另有一些研究改进了预训练模型中的表示。Kaneko 等[171]引入了一种微调技术以校正预训练上下文嵌入中的偏差。这种方法适用于各种预训练上下文嵌入模型, 无论是在标记级别还是句子级别, 而且无需重新对这些模型进行训练。Webster 等[172]调查了预训练语言模型在性别相关性方面的表现问题, 这种相关性可能对自然语言理解应用产生不必要的干扰, 如将性别与特定职业联系起来。研究者引入了一种相关性度量指标, 揭示了具有相似性能的模型可能以截然不同的速度对性别相关性进行编码, 提出了如何利用通用技术来减少测量到的性别相关性, 并突出了在解决问题时需要权衡不同策略的重要性。不同于以

往需要使用外部语料库进行微调的去偏差方法，Guo 等[173] 通过自动检测预训练语言模型中的偏差来解决问题。该研究采用了一种束搜索方法的变体用于自动搜索具有偏差的提示，使得模型在完形填空式的完成中对不同的人口群体表现出最大的差异。一旦有偏差的提示被确定，研究者引入了分布对齐损失来减轻模型的偏差。

4.2.3　虚假信息缓解措施

缓解大语言模型生成虚假信息的措施已经存在许多探索和研究，包括数据预处理、优化训练方式、改变解码方式、利用外部知识和模型编辑等。

数据预处理　大语言模型的知识主要在预训练阶段中获得[174]，所以预训练数据存在的矛盾数据或者错误数据就可能导致大语言模型幻觉。手动或者自动处理预训练数据来去除错误数据或者矛盾数据可以用于减轻幻觉现象。在传统自然语言生成任务上，已经存在了一系列的缓解幻觉的研究。可以根据源内容使用注释器从零开始编写忠实干净的数据[175]，但是这样的数据缺乏多样性[176]，不符合现实世界的复杂情况。另一种方法是使用注释器对网络现实数据进行改写[176]，如通过句法修改、去语境化等方式达到构建干净且忠实的数据。但是由于大语言模型的训练数据规模庞大，在训练前对数据进行构造、整理和修改变得越来越有挑战性。因此目前大语言模型通过自动选择可靠的数据或过滤掉噪声数据来达到构建干净忠实的数据集的目的。例如 GPT-3 通过与一系列高质量参考语料库的相似性计算进行数据清理，Falcon 的开发者则通过启发式规则从网络中精心提取高质量数据，并验证了正确管理相关语料库可以带来显著的性能提升。

优化训练方式　正如 Ranzato 等[177] 指出的，采用单词级别的最大似然估计训练可能导致暴露偏差，从而引发幻觉。相反，采用强化学习则可以缓解这一幻觉问题。而在强化学习中，奖励模型扮演着至关重要的角色。设计得当的奖励模型能够提供有效的训练信号，帮助模型实现减少幻觉的目标。举例来说，GPT-4 采用合成的幻觉数据进行训练，以奖励模型的方式，通过优化参数来减少模型的偏差。与 GPT-4 使用幻觉数据进行训练不同，John 等[178] 专门设计了一个特殊的奖励模型，其核心理念是鼓励大语言模型通过从专门设计的奖励中学习，挑战前提、表达不确定性并承认无知。这种方法被称为诚实导向的强化学习，它可以引导大语言模型探索自身的知识边界，使大模型能够拒绝回答超出自身能力范围的问题，而不是编造不真实的回答。

改变解码方式　3.3.2 节中提到提高多样性的解码策略，如 top-p 采样，与幻觉的增加呈正相关。鉴于此，Lee 等[129] 引入了一种被称为事实核采样的解码算法，旨在通过利用 top-p 和贪婪解码的优势，在多样性和事实性之间取得更有效的平衡。相应的，Li 等[179] 使用一种新颖的推理时间干预方法来提高大语言模型生成内容的真实性，缩小知道可信答案和输出可信答案之间的差距。与上述两种方法不同，Shi 等[180] 提出了一种简单的上下文感知解码策略，该方法旨在迫使大语言模型更多地关注上下文信息，而不是过度依赖自己的参数知识来做出决策。这项实验表明上下文感知解码策略有效地激发了大语言模型利用检索到的知识的能力，减少了对下游任务的事实幻觉。

利用外部知识　事实性问答等知识密集型任务对大语言模型的知识利用能力有很高的要求。然而由于下游领域相关训练数据缺乏、训练数据过时和模型幻觉等问题，大语言模型在相关任务中容易生成虚假信息[32]。为解决这一问题，大语言模型在知识密集型任务中往往通过构建外部知识库并在回答时检索相关知识条目，从而在不改变模型参数的前提下增强模型输出的正确率。现有的许多工作[181-183] 采用“先检索，再读取”的方法，即通过维护文本数据库的向量索引，并在推断时基于相似度检索相关数据以提高模型表现。与之相对的，“先生成，再检索”的方法[184] 则受人类完成任务时检查和验证过程的启发，首先使用大语言模型生成答案，再与检索所得的相关文字片段对比，并基于语言模型本身的能力进行评估和修正。而区别于上述基于文本知识库的两种方法，基于参数化的知识库方法使用另外训练的语言模型[185] 或者大语言模型自身生成任务的相关背景信息[186-187]，以减少虚假信息的产生。

模型编辑　模型编辑可以在一定数据范围中对模型进行编辑，以实现知识的插入、修改或删除，最终改变模型的输出。朴素的编辑方法是对模型进行微调，但效果并不理想[188]。目前主流的模型编辑方式分为两种，一种是冻结原模型并为其添加额外参数，另一种是修改模型内部参数。基于外部知识库的模型冻结方法通过为模型添加知识库和分类器，来保证模型优先使用知识库中的知识生成回答，而对模型本身参数没有改动[189]。相应的，在修改模型内部参数方面，也可以通过添加新的神经元帮助模型掌握新知识[190]，或基于元学习的思想预测模型的权重更改[191]。此外，基于模型存储信息位置的发现[192]，Meng 等[193] 使用定位知识后编辑的方法使编辑更加精准，并更有效地纠正模型错误输出。值得注意的是，现有修改模型内部参数的方法虽然能够提高模型在训练

数据上的表现, 但是也存在对知识的迁移能力不强[194]、在编辑多个知识后表现迅速下降[195]、对编辑区域外的输出存在影响[196] 等问题。

4.2.4 通用缓解措施

数据预处理 在大语言模型的数据收集阶段, 模型训练者可以采取一系列措施, 例如删除有害信息和敏感数据, 来对数据集进行清洗过滤, 以解决模型输出的安全性问题[69]。为了更有效地检测和自动移除问题片段, 训练者可以借助对生成内容的评估方法以及文本检测器, 如 BAD[197]、BBF[198]、DisSafety[199] 等。特别对于社交媒体数据集, 一种有效的策略是直接删除那些频繁发表有害言论的用户的所有信息, 以实现数据清理。此外, 还有一些研究致力于在解码层面避免模型的安全问题。例如, 使用 n 元阻塞或语句层级的 PPLM 可以直接将敏感或有害内容的抽样概率置为零, 从而有效防止这些内容被输入到大语言模型中[28]。通过这些综合的方法, 训练者可以在数据收集和模型解码阶段共同努力, 从而提高模型的安全性。

训练: 基于反馈的强化学习 有条件的文本生成模型 CTRL[200]、prefix-tuning[201]、扩散模型[202] 和强化学习等方法都在语言模型可控内容生成上取得了一定效果。而在大语言模型中, RLHF[34] 和其各类衍生方法成为了将模型价值观与人类对齐、控制模型内容输出的主要方法。

在 RLHF 中, 模型训练者会首先编写一个包含输入指令和期望输出的监督数据集。这些指令被输入到大语言模型中, 以生成一定数量的模型输出。随后标注人员会为输出打分或对生成的候选答案进行排序, 以反映人类偏好。最后模型训练者根据收集的人类偏好数据, 使用强化学习对大语言模型进行微调[26]。OpenAI 最先应用 RLHF 的思路对 GPT-3 进行微调, 得到了性能大幅提升的 InstructGPT[34]。相比有监督的指令微调, RLHF 有助于缓解模型幻觉和有毒信息的生成问题。

出于节约人工标注成本的需求并考虑到人工智能现有的成熟度, 一些研究者在 RLHF 的基础上提出了使用 AI 对大语言模型的输出内容生成反馈并微调的方法 RLAIF[51,203]。研究者为大语言模型设定了一些输出应该遵守的原则, 这被称为“宪法”, 并通过 fewshot 的方式引导大语言模型依据“宪法”对两个候选回答进行标注, 最后利用人工智能标注的数据集对模型进行强化学习训练。实验结果表明 RLAIF 和 RLHF 的效果在统计学上不具有显著性差异, 这说明了 RLAIF 的可行性。此外, 研究者还指出在强化学习之前可以先通过红队攻击让大语言模型输出有害内容, 并让模型依据“宪法”订正自己的答案, 最后将订正后的问答文本对用于对大语言模型的监督微调, 从而减少了强化学习阶段的模型训练时间[203]。

护栏 (Guardrails) Guardrails 不会改变模型参数, 而是通过在部署阶段使用检测器实时对模型的对话话题进行检测, 当涉及有害或敏感内容时模型会停止输出或结束聊天, 如“让我们换一个话题吧”[28,69]。从用户视角看, 用户输入可能涉及政治敏感、价值观错误、违法信息等内容, 或可能含有企图攻击模型的后门或指令, 检测器会识别出这些可能的情况并提前结束话题。从模型视角看, 利用 Guardrails 对模型的回答进行检测能直接屏蔽有害输出, 解决可能的信息侵害、有害信息生成问题, 并提升面临指令注入攻击时的安全性。这已经在 ChatGPT 上得到了验证[63]。

目前的研究者已经提出了大量针对话题的检测器。如 Roller 等[197] 介绍了一种涉及政治、种族、医疗、毒品的敏感话题分类检测器, 当上述话题被检测到时会触发模型的预设响应。此外借助仇恨言论驳斥数据集和外部知识, 大语言模型可以在检测到仇恨言论输入时对用户进行驳斥[69]。检测器还可以通过 OoD(out-of-distribution) 检测观察聊天内容是否与模型训练数据分布一致, 如果不一致直接拒绝回答, 以此增强模型回答的效果[28]。另外, 还有研究者提出了虚假信息检测[204]、代码生成质量检测、数学解析结果检测[28] 等, 但这些检测在大语言模型上的实际应用还有待进一步探索。而考虑到用户输入不受限制和场景的多样, 上述各类检测都应该具有更高的鲁棒性和泛化性, 以避免被越狱攻击等方式绕过, 同时需要注意过高的拒绝率会影响人机交互体验。

5 未来展望

现有研究对大语言模型的安全性认知还不够完整和全面, 并且随着大语言模型的发展, 大语言模型也会不断演化甚至产生新的安全性问题。本节将从安全性风险的评估、归因以及缓解措施 3 个方面讨论本文认为至关重要但目前研究还不够充分的问题。

5.1 安全性风险评估

评估方法是帮助研究人员发现大语言模型潜在问题的重要方法, 但是目前针对信息侵害、虚假信息、价值差异以及鲁棒性与可信度评估方法的研究还较为缺乏, 因此本文认为有必要持续性地发展这 4 个方面的评估方法。此外, 随着大语言模型的崛起, 对多模态大模型以及具身智能模型的研究也在日益增多。因此为了保证多模态大

模型与具身智能的安全发展，其安全性评估方法的研究刻不容缓。

信息侵害评估　诸如信息侵害等产生严重危害的大语言模型生成内容的评估方法研究目前还在初步阶段，据本文了解，目前关于大语言模型非法信息、隐私、敏感信息、版权的评估方法分为两种：基于提示回复和基于选择题，这两种方法是否能准确地评估大语言模型的信息侵害目前尚未可知。由于大语言模型的解码方式，其生成的内容随机性强，即使是同样的问题，大语言模型也可能会给出完全相反的回复。并且最近有研究表明大语言模型具有一定的场景感知 (situational awareness) 能力，能感知到其正在被测试，从而影响对它安全和对齐状况测试的结果[205]。因此研究一种或者几种能够更加准确、全面的信息侵害评估方法变得至关重要。

虚假信息的评估　虚假信息的评估不仅要包含忠实性、事实性的评估，还要包含过时信息的评估。过时信息的评估指的是对大语言模型参数中所学习到的信息进行过时性评估。目前的人工智能模型都不会自动地更新参数，这就意味着模型一旦训练完成就不会再有能力和知识上的进步，但实际上世界的信息知识是在不断更新的，甚至有些信息每分每秒都在更新。模型拥有的过时信息会导致生成错误的信息，因此对大语言模型的过时信息进行评估具有必要性，但是目前几乎没有这方面的研究工作。

价值差异评估　针对价值差异的评估方法目前停留在使用类似问卷调查的方法，将大语言模型作为一个人进行人格评估。这种评估方法只能通过向大语言模型提出一些特定的问题来调查大语言模型的回答以及反应，无法真正地模拟接近现实情境的复杂情况，所以模型可能在接受问卷调查的过程中表现良好，但是在面对真实情况时表现出错误的倾向。同时，生成式大语言模型的回答具有随机性，当用两种不同的描述询问同一个问题时，大语言模型的回答可能不同，这进一步加剧了使用问卷方式进行调查的不确定性，也体现了大语言模型回答的可操纵性，即研究人员可以通过设计提问方式来让大语言模型产生想要的输出。因此，需要设计一种更为客观准确的测量大语言模型价值观的方法。

鲁棒性与可信度的基准测试　相同的语义但是不同语法的输入可能导致大语言模型生成不同的结果，这表明大语言模型对于输入的理解并不鲁棒。虽然目前有一些关于鲁棒性评估的工作，但是仍然有很大的发展空间，例如包含更多的评估方面、更有效的评估方法。此外，有研究表明大语言模型的场景感知能力可能感知到其正处于评估测试中，从而影响模型输出内容，因此如何保证基准测试可信度至关重要。

多模态模型评估　多模态大模型是一种能够处理包括文本、图像、声音等多种感知模态的大型深度学习模型。这些模型结合了自然语言处理和计算机视觉等技术，能够理解和生成多种不同模态的数据并且支持跨模态信息的交互。但是，目前多模态大模型的评估也存在一些问题。一方面，随着 CLIP 等模型的不断发展，多模态大模型的应用产品不断增加，但是目前缺乏多模态大模型能力与安全性的评估方法，这使得在使用这些模型的过程中存在潜在的隐患。另一方面，多模态大模型同样可能遭受各类攻击。如 Carlini 等[19]成功设计对抗性样本对 MiniGPT-4、LLaVa 等多模态模型发起了指令注入攻击，并发现多模态模型比单模态模型更容易受到攻击。因此对于多模态模型开展安全性评估也十分必要。

具身智能安全性　具身智能模型可能存在物理上的安全性风险，因此建立完善的具身智能安全性评估至关重要。随着人工智能的不断发展，人们越来越关注如何让 AI 机器人像人类一样在真实世界中进行实践型学习。具身智能的含义并非仅仅指机器人本身，更涵盖了与环境交互以及在环境中实现特定功能的需求，这意味着，要让机器人像人类一样通过观察、移动、言语以及与世界的互动来学习，才能更好地适应复杂多变的环境，实现更高效的学习和应用[206]。具身智能模型往往以大语言模型为基础，并融入视觉、动作等其他模态信息[207]，因此大语言模型存在的安全性风险也可能存在于具身智能模型中。更进一步，当具身智能模型能够操控物理实体时，还可能带来更加严重的物理安全性风险。例如，机械臂失控对人类造成直接的伤害。因此，设计一套完善的具身智能模型安全性评估方法对于具身智能安全性发展具有重要意义。

5.2　安全性风险归因

本节深入探讨了大语言模型在安全性风险归因方面的两个关键方面：数据清洗和可解释性。这两个方面直接影响着模型的性能和可控性，对于解决大模型出现的安全性问题至关重要。

数据清洗　对训练数据的处理可以分为训练前清洗和训练后修正两个阶段。训练前的数据清洗已成为大语言模型训练的关键步骤之一。例如，Falcon 等[145]在训练前使用严格的数据清洗方法有效提高了模型性能。而在训练结束后，研究者同样可以检测并纠正由训练数据导致的问题。例如，Tanno 等[208]提出了一种框架，该框架通过

分析模型的错误输出, 找出数据集中导致这些错误的具体数据, 并针对性地消除这些数据对模型的影响。该方法不仅可以清洗训练数据, 还能在此基础上进一步提升模型性能。然而, 随着模型规模和训练数据量的持续增长, 以及越来越多的模型基于其他模型进行微调, 上述框架在当前大语言模型中的有效性仍待进一步研究。要解决训练数据带来的大模型安全性问题, 仍有大量工作需要研究, 如如何识别潜在的安全性风险数据, 如何在模型经过多次微调和对齐后准确找到问题数据, 以及清洗后的数据能在何种程度上改善模型的安全性等。这些问题的解决将有助于从根本上定位并解决大语言模型面临的安全性问题。

可解释性　迄今为止, 关于信息如何在深层神经网络中进行分发、组织和利用的问题仍然是大语言模型领域引人瞩目的谜团[26]。然而, 尽管现在仍然不清楚大模型内部各个神经元具体如何产生输出, 但是已经可以在一定程度上找到决定输出的关键神经元。例如, Meng 等[188] 提出了一种定位决定输出的关键神经元的方法, 实验证明定位并修改对应的神经元能够改变模型的输出。这些方向上的探索有利于了解模型的内在机制, 并为提高模型的安全性提供更具针对性的解决策略, 有助于建立更为可控的模型。此外, 大语言模型出现"涌现"现象的原因也是困扰研究人员的另一大谜团。尽管许多研究已经尝试解释为什么会产生这种突出表现[209], 但仍然缺乏正式的理论来深入剖析这一现象。

5.3　安全性风险缓解措施

在面对大语言模型存在的安全性问题时, 需要寻找有效的缓解措施以提高其应用的安全性和可信度。本节将探讨外部知识、模型编辑、大语言模型的回答边界、环境感知以及人类价值观评估解决措施等方面的策略。通过这些措施的综合运用, 有望在保持大语言模型高效性的同时, 最大程度地减轻其潜在的安全隐患, 为未来的应用提供更为可靠的基础。

外部知识　外部知识虽然可以缓解幻觉和过时知识等问题, 但是其仍然存在一些安全隐患, 具体而言, 外部知识的错误与过时信息会导致大语言模型的生成错误或过时的内容, 并且提供不相关的外部知识也可能会降低大语言模型生成内容的质量。因此, 如何构建忠实、干净并且能够实时更新的外部知识是未来需要探索的方向之一。

模型编辑　目前的模型编辑方法能够提高模型在训练数据上的表现, 但是也存在对知识的迁移能力不强、在编辑多个知识后表现会迅速下降、对编辑区域外的输出造成影响等问题。另外, 随着大语言模型参数量的逐步提高, 原有的方法是否仍然有效仍需研究。另外, 如何在计算资源受限时进行模型编辑、如何对闭源大语言模型进行编辑等问题也需要进一步探索。

大语言模型的回答边界　目前开放使用的大语言模型 (例如 ChatGPT、文心一言、讯飞星火等) 为了避免出现生成内容的问题, 都对大语言模型的回复做出了边界限制, 针对敏感问题采用了模糊回复或者拒绝回复。这样的解决策略确实能缓解一些安全性问题, 但是却也降低了大语言模型的回复能力, 因此如何能够在不降低大语言模型的回复能力的前提下提高安全性是未来需要研究的方向。

环境感知　环境感知 (context awareness) 指模型通过其他设备或用户对周围环境进行参数采集、语义表达、语义查询解析和语义推理的能力。传统的环境感知已经在推荐系统、无人驾驶、物联网等领域得到了广泛应用。而随着大语言模型和人工智能 Agent 的出现, 人工智能可以通过调用外部设备接口获取周围环境信息, 以此做出更具针对性的决策, 例如通过获取所在国家、地区的价值观和法规信息来提高生成内容的合法性和合规性。从应用上看, Wang 等[210] 在针对游戏的人工智能 Agent 中利用环境感知帮助大模型获取游戏环境状态并做出针对性决策, 但环境感知在现实场景下的人工智能 Agent 应用还有待进一步研究。

人类价值观评估解决措施　研究人员需要提出新的测评大语言模型价值观的方法, 新方法应以更稳定的、更接近真实情境的方式测试大语言模型中蕴含不同人类价值观的程度。例如, 多层面价值观评估: 将人类价值观分为不同层面, 如道德、文化、政治和伦理等, 以更全面地了解模型的性能。这样可以确保模型在各种领域中都能正确地反映出不同层面的价值观。情境敏感性测试: 模型应在多种真实世界情境下进行测试, 以确保其在不同情境下的表现稳定。在医疗、法律、商业等不同领域中测试模型, 以评估其对特定情境的适应能力。多源数据集集成: 利用多源数据集, 包括文本、图像、音频和视频等, 来评估模型的价值观。这有助于模型更好地理解人类价值观的多样性, 而不仅仅是从文本数据中学习。

6　结束语

随着大语言模型的迅猛发展, 其在众多领域展现出强大的能力, 但同时引发的安全性问题也日益凸显。如图 3 所示, 本文从大语言模型自身

安全风险和生成内容安全风险两方面进行了全面的总结，并调研了相应的评估方法。然后，针对评估发现的问题，本文深入分析了潜在的原因。更进一步，本文还回顾了目前针对大语言模型安全性问题的缓解措施，旨在将复杂的安全问题分解为更小的问题，以逐步解决。最后，本文还对大语言模型安全性相关的研究进行了展望，希望能为该领域的研究提供启发和指导。

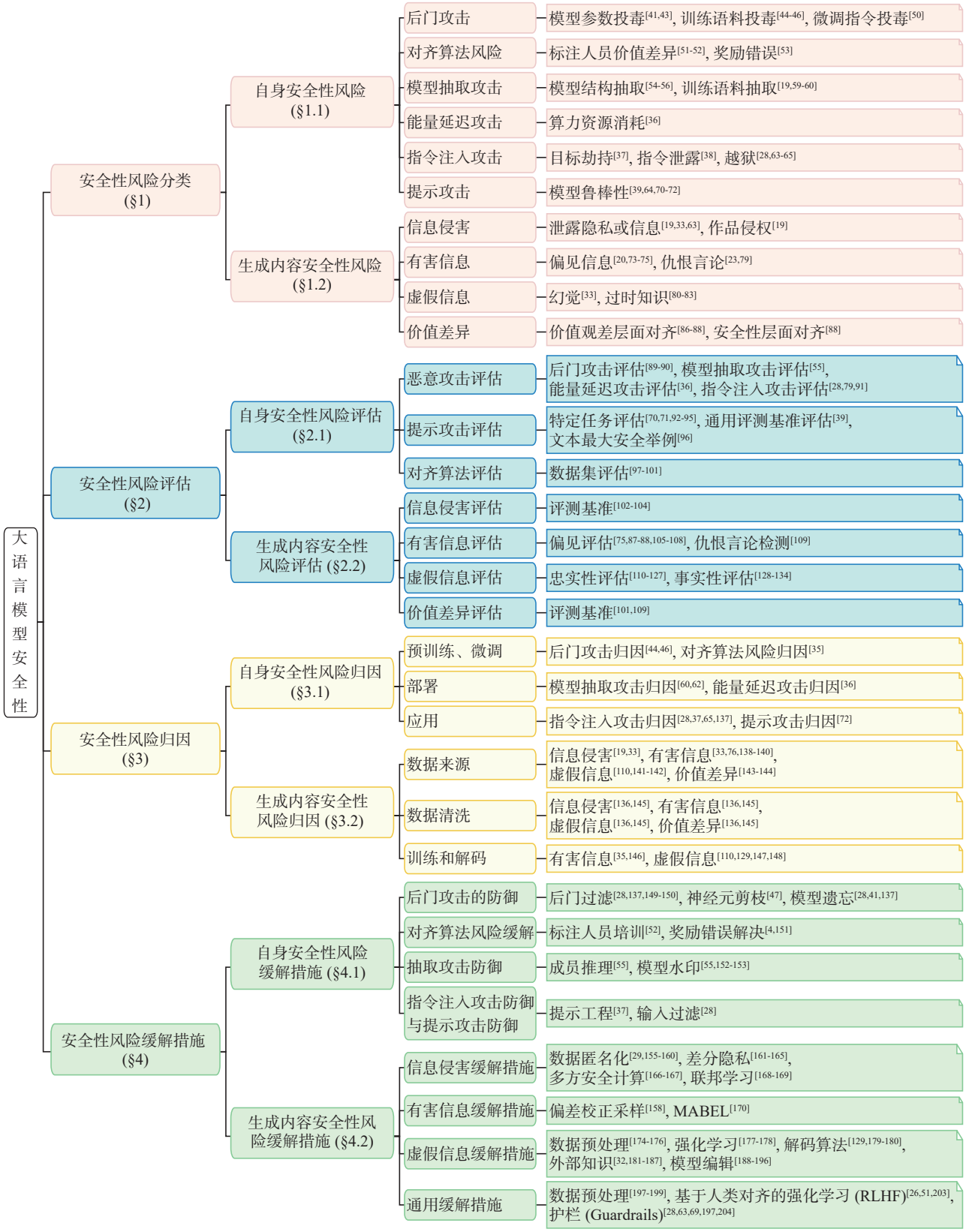

图 3　文章结构示意

Fig. 3　Structure diagram of the article

致谢　李思霖、兰天伟、邱昱力与柳泽明作者对本文的贡献度相同。感谢国家自然基金与科技创新 2030—“新一代人工智能”重大项目的资助与支持!

参考文献:

[1] WEI J, TAY Y, BOMMASANI R, et al. Emergent abilities of large language models[EB/OL]. (2022−07−15)[2024−01−03]. https://arxiv.org/abs/2206.07682.

[2] LIU Xiao, JI Kaixuan, FU Yicheng, et al. P-Tuning: prompt tuning can be comparable to fine-tuning across scales and tasks[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin: Association for Computational Linguistics. 2022: 61–68.

[3] HAO Jianye, YANG Tianpei, TANG Hongyao, et al. Exploration in deep reinforcement learning: from single-agent to multiagent domain[EB/OL]. (2021−09−14)[2024−01−03]. https://arxiv.org/abs/2109.06668v6.

[4] JI Jiamin, QIU Tianyi, CHEN Boyuan, et al. AI alignment: a comprehensive survey[EB/OL]. (2023−10−30)[2024−01−03]. https://arxiv.org/abs/2310.19852.

[5] 鲍小异, 姜晓彤, 王中卿, 等. 基于跨语言图神经网络模型的属性级情感分类[J]. 软件学报, 2023, 34(2): 676−689.
BAO Xiaoyi, JIANG Xiaotong, WANG Zhongqing, et al. Cross-lingual aspect-level sentiment classification with graph neural network[J]. Journal of software, 2023, 34(2): 676−689.

[6] EFRON, B, TIBSHIRANI R, JEROME F. The elements of statistical learning[M]. London: Springer, 2009.

[7] DENG Li, YU Dong. Deep learning: methods and applications[J]. Foundations and trends® in signal processing, 2014, 7(3−4): 197−387.

[8] BUBECK S, CHANDRASEKARAN V, ELDAN R, et al. Sparks of artificial general intelligence: early experiments with GPT-4[EB/OL]. (2023−03−22)[2024−01−03]. https://arxiv.org/abs/2303.12712v5.

[9] OpenAI. GPT-4 technical report[EB/OL]. (2023−08−30)[2024−01−03]. https://api.semanticscholar.org/CorpusID:266362871.

[10] DONG Qingxiu, LI Lei, DAI Damai, et al. A survey on in-context learning[EB/OL]. (2022−12−31)[2024−01−03]. https://arxiv.org/abs/2301.00234.

[11] LIU Pengfei, YUAN Weizhe, FU Jinlan, et al. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing[J]. ACM computing surveys, 2023, 55(9): 1−35.

[12] WEI J, WANG Xuezhi, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models[EB/OL]. (2022−01−28)[2024−01−03]. https://arxiv.org/abs/2201.11903v6.

[13] SON Guijin, JUNG Hanna, JIN S. Beyond classification: financial reasoning in state-of-the-art language models[EB/OL]. (2023−05−30)[2024−01−03]. https://api.semanticscholar.org/CorpusID:258437058.

[14] BLAIR-STANEK A, HOLZENBERGER N, VAN DURME B. Can GPT-3 perform statutory reasoning?[C]// Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law. Braga Portugal: ACM, 2023: 22−31.

[15] YU Fang, QUARTEY L, SCHILDER F. Legal prompting: teaching a language model to think like a lawyer [EB/OL]. (2022−12−30)[2024−01−03]. https://api.semanticscholar.org/CorpusID:254221002.

[16] TANG Ruixiang, HAN Xiaotian, JIANG Xiaoqian, et al. Does synthetic data generation of LLMs help clinical text mining?[EB/OL]. (2023−03−25)[2024−01−03]. https://api.semanticscholar.org/CorpusID:257405132.

[17] ALEC R, JEFFREY W, REWON C, et al. Language models are unsupervised multitask learners[EB/OL]. [2024−01−03]. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[18] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc. , 2020: 1877−1901.

[19] CARLINI N, TRAMER F, WALLACE E, et al. Extracting training data from large language models[C]//30th USENIX Security Symposium (USENIX Security 21). [S. l.]: [s. n.], 2021: 2633−2650.

[20] ABID A, FAROOQI M, ZOU J. Persistent anti-muslim bias in large language models[C]//Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. Virtual Event: ACM, 2021: 298−306.

[21] TAYLOR R, KARDAS M, CUCURULL G, et al. Galactica: a large language model for science[EB/OL]. (2022−11−16)[2024−01−03]. https://arxiv.org/abs/2211.09085v1.

[22] EDWARDS B. New meta AI demo writes racist and inaccurate scientific literature, gets pulled[EB/OL]. (2022−11−18)[2023−09−27]. https://arstechnica.com/information-technology/2022/11/after-controversy-meta-pulls-demo-of-ai-model-that-writes-scientific-papers/.

[23] RAE J W, BORGEAUD S, CAI T, et al. Scaling language models: methods, analysis & insights from training gopher[EB/OL]. (2021−12−08)[2024−01−03]. https://arxiv.org/abs/2112.11446v2.

[24] 任奎, 孟泉润, 闫守琨, 等. 人工智能模型数据泄露的攻击与防御研究综述[J]. 网络与信息安全学报, 2021, 7(1): 1−10.
REN Kui, MENG Quanrun, YAN Shoukun, et al. Survey of artificial intelligence data security and privacy protection[J]. Chinese journal of network and information security, 2021, 7(1): 1−10.

[25] GOLDSTEIN J A, SASTRY G, MUSSER M, et al. Generative language models and automated influence operations: emerging threats and potential mitigations [EB/OL]. (2023−01−10)[2024−01−03]. https://arxiv.org/abs/2301.04246v1.

[26] ZHAO W X, ZHOU Kun, LI Junyi, et al. A survey of large language models[EB/OL]. (2023−03−31)[2024−01−03]. https://arxiv.org/abs/2303.18223v15.

[27] WANG Yufei, ZHONG Wanjun, LI Liangyou, et al. Aligning large language models with human: a survey [EB/OL]. (2023−07−28)[2024−01−03]. https://api.semanticscholar.org/CorpusID:260356605.

[28] HUANG Xiaowei, RUAN Wenjie, HUANG Wei, 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.

[29] LIU Yang, YAO Yuanshun, TON J F, et al. Trustworthy LLMs: a survey and guideline for evaluating large language models' alignment[EB/OL]. (2023−08−10)[2024−01−03]. https://arxiv.org/abs/2308.05374.

[30] ZHANG Yue, LI Yafu, CUI Leyang, et al. Siren's song in the AI ocean: a survey on hallucination in large language models[EB/OL]. (2023−09−03)[2024−01−03]. https://arxiv.org/abs/2309.01219.

[31] RAWTE V, SHETH A, DAS A. A survey of hallucination in large foundation models[EB/OL]. (2023−09−12)[2024−01−03]. https://arxiv.org/abs/2309.05922.

[32] CHAGN Yupeng, WANG Xu, WANG Jindong, et al. A survey on evaluation of large language models[EB/OL]. (2023−07−06)[2024−01−03]. https://arxiv.org/abs/2307.03109?context=cs.AI.

[33] WEIDINGER L, UESATO J, RAUH M, et al. Taxonomy of risks posed by language models[C]//2022 ACM Conference on Fairness, Accountability, and Transparency. Seoul: ACM, 2022: 214−229.

[34] OUYANG Long, WU J, JIANG Xu, et al. Training language models to follow instructions with human feedback[EB/OL]. (2022−03−04)[2024−01−03]. https://arxiv.org/abs/2203.02155?context=cs.CL.

[35] STIENNON N, OUYANG Long, WU J, et al. Learning to summarize from human feedback[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc. , 2020: 3008−3021.

[36] SHUMAILOV I, ZHAO Yiren, BATES D, et al. Sponge examples: energy-latency attacks on neural networks[C]//2021 IEEE European Symposium on Security and Privacy. Vienna: IEEE, 2021: 212−231.

[37] GRESHAKE K, ABDELNABI S, MISHRA S, et al. Not what you've signed up for: compromising real-world LLM-integrated applications with indirect prompt injection[C]//Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security. Copenhagen: ACM, 2023: 79−90.

[38] PEREZ F, RIBEIRO I. Ignore previous prompt: attack techniques for language models[EB/OL]. (2022−11−18)[2024−01−03]. https://api.semanticscholar.org/CorpusID:253581710.

[39] ZHU Kaijie, WANG Jindong, ZHOU Jiaheng, et al. PromptRobust: towards evaluating the robustness of large language models on adversarial prompts[EB/OL]. (2023−06−07) [2024−01−03]. https://arxiv.org/abs/2306.04528.

[40] 冀甜甜, 方滨兴, 崔翔, 等. 深度学习赋能的恶意代码攻防研究进展[J]. 计算机学报, 2021, 44(4): 669−695.
JI Tiantian, FANG Binxing, CUI Xiang, et al. Research on deep learning-powered malware attack and defense techniques[J]. Chinese journal of computers, 2021, 44(4): 669−695.

[41] LI Linyang, SONG Demin, LI Xiaonan, et al. Backdoor attacks on pre-trained models by layerwise weight poisoning[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic. Stroudsburg: Association for Computational Linguistics, 2021: 3023−3032.

[42] DEVLIN J, CHANG Mingwei, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: Association for Computational Linguistics, 2019: 4171−4186.

[43] YANG Wenkai, LI Lei, ZHANG Zhiyuan, et al. Be careful about poisoned word embeddings: exploring the vulnerability of the embedding layers in NLP models[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2021: 2048−2058.

[44] LI Shaofeng, LIU Hui, DONG Tian, et al. Hidden backdoors in human-centric language models[C]//Proceed-

ings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. Virtual Event: ACM, 2021: 3123−3140.

[45] CHEN Kangjie, MENG Yuxian, SUN Xiaofei, et al. BadPre: task-agnostic backdoor attacks to pre-trained NLP foundation models[C]//International Conference on Learning Representations. Virtual Event: ICLR, 2021.

[46] CHEN Xiaoyi, SALEM A, CHEN Dingfan, et al. BadNL: backdoor attacks against NLP models with semantic-preserving improvements[EB/OL]. (2020−06−06)[2024−01−03]. https://arxiv.org/abs/2006.01043v2.

[47] ZHANG Zhengyan, XIAO Guangxuan, LI Yongwei, et al. Red alarm for pre-trained models: universal vulnerability to neuron-level backdoor attacks[J]. Machine intelligence research, 2023, 20(2): 180−193.

[48] LIU Yinhan, OTT M, GOYAL N, et al. RoBERTa: a robustly optimized BERT pretraining approach[EB/OL]. (2019−07−06)[2024−01−03]. https://arxiv.org/abs/1907.11692v1.

[49] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale[EB/OL]. (2020−10−22)[2024−01−03]. https://arxiv.org/abs/2010.11929.

[50] XU Jiashu, MA M D, WANG Fei, et al. Instructions as backdoors: backdoor vulnerabilities of instruction tuning for large language models[EB/OL]. (2023−05−24)[2024−01−03]. https://arxiv.org/abs/2305.14710.

[51] LEE H, PHATALE S, MANSOOR H, et al. RLAIF vs. RLHF: scaling reinforcement learning from human feedback with AI feedback[EB/OL]. (2023−09−01)[2024−01−03]. https://arxiv.org/abs/2309.00267.

[52] KIRK H R, VIDGEN B, RÖTTGER P, et al. Personalisation within bounds: a risk taxonomy and policy framework for the alignment of large language models with personalised feedback[EB/OL]. (2023−03−10)[2024−01−03]. https://arxiv.org/abs/2303.05453.

[53] PAN A, BHATIA K, STEINHARDT J. The effects of reward misspecification: mapping and mitigating misaligned models[EB/OL]. (2022−01−10)[2024−01−03]. https://arxiv.org/abs/2201.03544.

[54] OREKONDY T, SCHIELE B, FRITZ M. Knockoff nets: stealing functionality of black-box models[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 4949−4958.

[55] KRISHNA K, TOMAR G S, PARIKH A P, et al. Thieves on sesame street! model extraction of BERT-based APIs[EB/OL]. (2019−10−06)[2024−01−03]. https://arxiv.org/abs/1910.12366v3.

[56] LIU Yupei, JIA Jinyuan, LIU Hongbin, et al. StolenEncoder: stealing pre-trained encoders in self-supervised learning[C]//Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security. Los Angeles: ACM, 2022: 2115−2128.

[57] DENG Jia, DONG Wei, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE, 2009: 248−255.

[58] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//Proceedings of the 38th International Conference on Machine Learning. Virtual Event: PMLR, 2021: 8748−8763.

[59] ELMAHDY A, SALEM A. Deconstructing classifiers: towards a data reconstruction attack against text classification models[C]//Proceedings of the Fifth Workshop on Privacy in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2024: 143–158.

[60] ALI A K, MALIHEH L, ARIE van D. Targeted attack on GPT-Neo for the SATML language model data extraction challenge[EB/OL]. (2023−02−13)[2024−01−03]. https://arxiv.org/abs/2302.07735.

[61] BLACK S, BIDERMAN S, HALLAHAN E, et al. GPT-NeoX-20B: an open-source autoregressive language model[EB/OL]. (2022−06−14) [2024−01−03]. https://arxiv.org/abs/2204.06745v1.

[62] ZANELLA-BÉGUELIN S, WUTSCHITZ L, TOPLE S, et al. Analyzing information leakage of updates to natural language models[C]//Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security. Virtual Event: ACM, 2020: 363−375.

[63] LI Haoran, GUO Dadi, FAN Wei, et al. Multi-step jailbreaking privacy attacks on ChatGPT[EB/OL]. (2023−04−11)[2024−01−03]. https://arxiv.org/abs/2304.05197.

[64] CHEN Lingjiao, ZAHARIA M, ZOU J. How is ChatGPT's behavior changing over time?[J/OL]. Harvard data science review, (2024−03−13)[2024−11−15]. https://doi.org/10.1162/99608f92.5317da47.

[65] ZOU A, WANG Zifan, NICHOLAS C, et al. Universal and transferable adversarial attacks on aligned language models[EB/OL]. (2023−07−17)[2024−01−03]. https://arxiv.org/abs/2307.15043.

[66] YUAN Youliang, JIAO Wenxiang, WANG Wenxuan, et al. GPT-4 is too smart to be safe: stealthy Chat with LLMs via cipher[EB/OL]. (2023−08−12)[2024−01−03]. https://arxiv.org/abs/2308.06463?context=cs.

[67] SHEN Xinyue, CHEN Z, BACKES M, et al. In ChatGPT we trust? measuring and characterizing the reliability of ChatGPT[EB/OL]. (2023−04−18)[2024−01−03].

https://api.semanticscholar.org/CorpusID:258187122.

[68] WANG Jindong, HU Xixu, HOU Wenxin, et al. On the robustness of ChatGPT: an adversarial and out-of-distribution perspective[C]//ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models. [S. l.]: ICLR, 2023: 48−62.

[69] DEGN Jiawen, CHENG Jiale, SUN Hao, et al. Towards safer generative language models: a survey on safety risks, evaluations, and improvements [EB/OL]. (2023−11−30)[2024−01−03]. https://arxiv.org/abs/2302.09270v3.

[70] HUGGING Face. 为大语言模型建立红队对抗[EB/OL]. (2023−02−24)[2023−09−27]. https://huggingface.co/blog/zh/red-teaming.

[71] BORKAR J. What can we learn from data leakage and unlearning for law?[EB/OL]. (2023−07−19)[2024−01−03]. https://arxiv.org/abs/2307.10476.

[72] KIM S, YUN S, LEE H, et al. ProPILE: probing privacy leakage in large language models[EB/OL]. (2023−07−04)[2024−01−03]. https://arxiv.org/abs/2307.01881.

[73] BOSTROM N. Information hazards: a typology of potential harms from knowledge[J]. Review of contemporary philosophy, 2011(10): 44−79.

[74] COLIN R, NOAM S, ADAM R, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. Journal of machine learning research, 2020, 21(1): 140: 5485−140: 5551.

[75] SCHRAMOWSKI P, TURAN C, ANDERSEN N, et al. Large pre-trained language models contain human-like biases of what is right and wrong to do[J]. Nature machine intelligence, 2022, 4: 258−268.

[76] CALISKAN A, BRYSON J J, NARAYANAN A. Semantics derived automatically from language corpora contain human-like biases[J]. Science, 2017, 356(6334): 183−186.

[77] LUCY L, BAMMAN D. Gender and representation bias in GPT-3 generated stories[C]//Proceedings of the Third Workshop on Narrative Understanding. Stroudsburg: Association for Computational Linguistics, 2021: 48−55.

[78] HARTMANN J, SCHWENZOW J, WITTE M. The political ideology of conversational AI: converging evidence on ChatGPT's pro-environmental, left-libertarian orientation[J]. SSRN electronic journal, 2023: 4216084.

[79] GEHMAN S, GURURANGAN S, SAP M, et al. RealToxicityPrompts: evaluating neural toxic degeneration in language models[C]//Findings of the Association for Computational Linguistics: EMNLP 2020. Stroudsburg: Association for Computational Linguistics, 2020: 3356−3369.

[80] 汪楠、成鹰、曹辉, 等. 信息检索技术[M]. 第 3 版. 北京: 清华大学出版社, 2018.
WANG Nan, CHENG Ying, CAO Hui, et al. Information Retrieval Technology[M]. Third edition. Beijing: Tsinghua University Press, 2018.

[81] FU Xiaorong, ZHANG Bin, et al. Impact of quantity and timeliness of EWOM information on consumer's online purchase intention under C2C environment[J]. Asian journal of business research, 2011, 1(2): 110010.

[82] EDMONDS C T, EDMONDS J E, VERMEER B Y, et al. Does timeliness of financial information matter in the governmental sector?[J]. Journal of accounting and public policy, 2017, 36(2): 163−176.

[83] OUYANG Long, WU J, XU Jiang, et al. Training language models to follow instructions with human feedback[EB/OL]. (2022−05−04)[2024−01−03]. https://arxiv.org/abs/2203.02155.

[84] ROKEACH M. The nature of human values[M]. New York: The Free Press, 1973.

[85] DIGNUM V. Responsible artificial intelligence: how to develop and use AI in a responsible way[M]. Cham: Springer International Publishing, 2019

[86] HÄMMERL K, DEISEROTH B, SCHRAMOWSKI P, et al. Do multilingual language models capture differing moral norms?[EB/OL]. (2022−03−18)[2024−01−03]. https://arxiv.org/abs/2203.09904.

[87] TOUILEB S, ØVRELID L, VELLDAL E. Occupational biases in Norwegian and multilingual language models[C]//Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing. Seattle: Association for Computational Linguistics, 2022: 200−211.

[88] HAEMMERL K, DEISEROTH B, SCHRAMOWSKI P, et al. Speaking multiple languages affects the moral bias of language models[C]//Findings of the Association for Computational Linguistics: ACL 2023. Toronto: Association for Computational Linguistics, 2023: 2137−2156.

[89] WU Baoyuan, CHEN Hongrui, ZHANG Mingda, et al. BackdoorBench: a comprehensive benchmark of backdoor learning[EB/OL]. (2022−06−25)[2024−01−03]. https://arxiv.org/abs/2206.12654.

[90] SHENG Xuan, HAN Zhaoyang, LI Piji, et al. A survey on backdoor attack and defense in natural language processing[C]//2022 IEEE 22nd International Conference on Software Quality, Reliability and Security. Guangzhou: IEEE, 2022: 809−820.

[91] XU Jing, JU Da, LI M, et al. Bot-adversarial dialogue for safe conversational agents[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2021: 2950−2968.

[92] JIAO Wenxiang, WANG Wenxuan, HUANG J T, et al. Is ChatGPT a good translator? yes with GPT-4 as the engine[EB/OL]. (2023−01−20)[2024−01−03]. https://api.semanticscholar.org/CorpusID:257631519.

[93] CHEN Shan, KANN B, FOOTE M, et al. The utility of ChatGPT for cancer treatment information[EB/OL] (2023−03−23)[2024−01−03]. https://api.semanticscholar.org/CorpusID:257686179.

[94] WANG Boxin, XU Chejian, WANG Shuohang, et al. Adversarial GLUE: a multi-task benchmark for robustness evaluation of language models[C]//Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2). [S. l.]: NeurIPS, 2021.

[95] NIE Yixin, WILLIAMS A, DINAN E, et al. Adversarial NLI: a new benchmark for natural language understanding[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020: 4885−4901.

[96] LA MALFA E, WU Min, LAURENTI L, et al. Assessing robustness of text classification through maximal safe radius computation[C]//Findings of the Association for Computational Linguistics: EMNLP 2020. Stroudsburg: Association for Computational Linguistics, 2020: 2949−2968.

[97] PARRISH A, CHEN A, NANGIA N, et al. BBQ: a hand-built bias benchmark for question answering[C]// Findings of the Association for Computational Linguistics: ACL 2022. Dublin: Association for Computational Linguistics, 2022: 2086−2105.

[98] ZHAO Jieyu, WANG Tianlu, YATSKAR M, et al. Gender bias in coreference resolution: evaluation and debiasing methods[C]//Proceedings of the 2018 Conference of the North American Chapter ofthe Association for Computational Linguistics: Human LanguageTechnologies, Volume 2 (Short Papers). New Orleans: Association for Computational Linguistics, 2018: 15–20.

[99] NANGIA N, VANIA C, BHALERAO R, et al. CrowS-pairs: a challenge dataset for measuring social biases in masked language models[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2020: 1953−1967.

[100] ROSENTHAL S, ATANASOVA P, KARADZHOV G, et al. SOLID: a large-scale semi-supervised dataset for offensive language identification[EB/OL]. (2020−04−29) [2024−01−03]. https://arxiv.org/abs/2004.14454.

[101] SORENSEN T, JIANG Liwei, HWANG J D, et al. Value kaleidoscope: engaging AI with pluralistic human values, rights, and duties[EB/OL]. (2023−09−02) [2024−01−03]. https://api.semanticscholar.org/CorpusID:261531157.

[102] SUN Hao, ZHANG Zhexin, DENG Jiawen, et al. Safety assessment of chinese large language models[EB/OL]. (2023−04−20)[2024−01−03]. https://arxiv.org/abs/2304.10436.

[103] XU Guohai, LIU Jiayi, YAN Ming, et al. CValues: measuring the values of Chinese large language models from safety to responsibility[EB/OL]. (2023−07−19) [2024−01−03]. https://arxiv.org/abs/2307.09705.

[104] ZHANG Zhexin, LEI Leqi, WU Lindong, et al. SafetyBench: evaluating the safety of large language models with multiple choice questions[EB/OL]. (2023−09−13) [2024−01−03]. https://arxiv.org/abs/2309.07045.

[105] NOZZA D, BIANCHI F, LAUSCHER A, et al. Measuring harmful sentence completion in language models for LGBTQIA+ individuals[C]//Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion. Dublin: Association for Computational Linguistics, 2022: 26−34.

[106] FAISAL F, ANASTASOPOULOS A. Geographic and geopolitical biases of language models[EB/OL]. (2022−12−20)[2024−01−03]. https://api.semanticscholar.org/CorpusID:254877109.

[107] SIMMONS G. Moral mimicry: large language models produce moral rationalizations tailored to political identity[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop). Toronto: Association for Computational Linguistics, 2023: 282−297.

[108] ZHANG Susan, ROLLER S, GOYAL N, et al. OPT: open pre-trained transformer language models[EB/OL]. (2022−05−02)[2024−01−03]. https://api.semanticscholar.org/CorpusID:248496292.

[109] FENG Shangbin, PARK C Y, LIU Yuhan, et al. From pretraining data to language models to downstream tasks: tracking the trails of political biases leading to unfair NLP models[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto: Association for Computational Linguistics, 2023: 11737−11762.

[110] JI Ziwei, LEE N, FRIESKE R, et al. Survey of hallucination in natural language generation[J]. ACM computing surveys, 2023, 55(12): 1−38.

[111] SHUSTER K, POFF S, CHEN Moya, et al. Retrieval augmentation reduces hallucination in conversation[C]// Findings of the Association for Computational Linguistics: EMNLP 2021. Stroudsburg: Association for Computational Linguistics, 2021: 3784−3803.

[112] SUN Yanli. Mining the correlation between human and automatic evaluation at sentence level[EB/OL]. (2010−06−28)[2024−01−03]. http://www.lrec-conf.org/proceedings/lrec2010/pdf/87_Paper.pdf.

[113] DHINGRA B, FARUQUI M, PARIKH A, et al. Handling divergent reference texts when evaluating table-to-text generation[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019: 4884−4895.

[114] WANG Zhenyi, WANG Xiaoyang, AN Bang, et al. Towards faithful neural table-to-text generation with content-matching constraints[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020: 1072−1086.

[115] GOODRICH B, RAO V, LIU P J, et al. Assessing the factual accuracy of generated text[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Anchorage: ACM, 2019: 166−175.

[116] DURMUS E, HE He, DIAB M. FEQA: a question answering evaluation framework for faithfulness assessment in abstractive summarization[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020: 5055−5070.

[117] HONOVICH O, CHOSHEN L, AHARONI R, et al. Q2: evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2021: 7856−7870.

[118] REBUFFEL C, SCIALOM T, SOULIER L, et al. DataQuestEval: a referenceless metric for data-to-text semantic evaluation[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2021: 8029−8036.

[119] DUŠEK O, KASNER Z. Evaluating semantic accuracy of data-to-text generation with natural language inference[C]//Proceedings of the 13th International Conference on Natural Language Generation. Dublin: Association for Computational Linguistics, 2020: 131−137.

[120] DZIRI N, RASHKIN H, LINZEN T, et al. Evaluating attribution in dialogue systems: the BEGIN benchmark[J]. Transactions of the association for computational linguistics, 2022, 10: 1066−1083.

[121] FALKE T, RIBEIRO L F R, UTAMA P A, et al. Ranking generated summaries by correctness: an interesting but challenging application for natural language inference[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019: 2214−2220.

[122] HUANG Yichong, FENG Xiachong, FENG Xiaocheng, et al. The factual inconsistency problem in abstractive text summarization: a survey[EB/OL]. (2021−04−30)[2024−01−03]. https://arxiv.org/abs/2104.14839.

[123] DINAN E, ROLLER S, SHUSTER K, et al. Wizard of wikipedia: knowledge-powered conversational agents [EB/OL]. (2018−11−03)[2024−01−03]. https://arxiv.org/abs/1811.01241v2.

[124] LIU Tianyu, ZHANG Yizhe, BROCKETT C, et al. A token-level reference-free hallucination detection benchmark for free-form text generation[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin: Association for Computational Linguistics, 2022: 6723−6737.

[125] SASHANK S, BEHNAM H, SPANDANA, et al. Rome was built in 1776: a case study on factual correctness in knowledge-grounded response generation[EB/OL]. (2021−10−11) [2024−01−03]. https://arxiv.org/abs/2110.05456.

[126] ZHOU Chunting, NEUBIG G, GU Jiatao, et al. Detecting hallucinated content in conditional neural sequence generation[C]//Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Stroudsburg: Association for Computational Linguistics, 2021: 1393−1404.

[127] GOYAL T, DURRETT G. Evaluating factuality in generation with dependency-level entailment[C]//Findings of the Association for Computational Linguistics: EMNLP 2020. Stroudsburg: Association for Computational Linguistics, 2020: 3592−3603.

[128] LIN S, HILTON J, EVANS O. TruthfulQA: measuring how models mimic human falsehoods[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin: Association for Computational Linguistics, 2022: 3214−3252.

[129] LEE N, WEI Ping, PENG Xu, et al. Factuality enhanced language models for open-ended text generation[C]//Advances in Neural Information Processing Systems. New Orleans: NeurIPS, 2022: 34586−34599.

[130] MIN S, KRISHNA K, LYU Xinxi, et al. FActScore: fine-grained atomic evaluation of factual precision in long form text generation[C]//Proceedings of the 2023

Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2023: 10523436.

[131] LI Junyi, CHENG Xiaoxue, ZHAO Xin, et al. HaluEval: a large-scale hallucination evaluation benchmark for large language models[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics, 2023: 6449–6464.

[132] YU Jifan, WANG Xiaozhi, TU Shangqing, et al. KoLA: carefully benchmarking world knowledge of large language models[EB/OL]. (2023−06−16)[2024−01−03]. https://arxiv.org/abs/2306.09296.

[133] LIN C Y. ROUGE: A package for automatic evaluation of summaries[C]//Text Summarization Branches Out. Spain: Association for Computational Linguistics, 2004: 74−81.

[134] ZHA Yuheng, YANG Yichi, LI Ruichen, et al. AlignScore: evaluating factual consistency with A unified alignment function[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto: Association for Computational Linguistics, 2023: 11328−11348.

[135] NI Jianmo, QU Chen, LU Jing, et al. Large dual encoders are generalizable retrievers[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2022: 9844−9855.

[136] TOUVRON H, LAFRIL T, GAUTIER I, et al. LLaMA: open and efficient foundation language models[EB/OL]. (2023−02−27)[2024−01−03]. https://arxiv.org/abs/2302.13971.

[137] GRESHAKE K, ABDELNABI S, MISHRA S, et al. More than you've asked for: a comprehensive analysis of novel prompt injection threats to application-integrated large language models[EB/OL]. (2023−02−23)[2024−01−03]. https://arxiv.org/abs/2302.12173v1.

[138] HOLSTEIN K, WORTMAN VAUGHAN J, DAUMÉ H III, et al. Improving fairness in machine learning systems: what do industry practitioners need? [C]//Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. Glasgow: ACM, 2019: 1−16.

[139] SAMBASIVAN N, ARNESEN E, HUTCHINSON B, et al. Re-imagining algorithmic fairness in India and beyond[C]//Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. Virtual Event: ACM, 2021: 315−328.

[140] FORTUNA P, NUNES S. A survey on automatic detection of hate speech in text[J]. ACM computing surveys, 2019, 51(4): 1−30.

[141] MCKENNA N, LI Tianyi, CHENG Liang, et al. Sources of hallucination by large language models on inference tasks[EB/OL]. (2023−05−24)[2024−01−03]. https://arxiv.org/abs/2305.14552.

[142] WISEMAN S, SHIEBER S, RUSH A. Challenges in data-to-document generation[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: Association for Computational Linguistics, 2017: 2253−2263.

[143] SCHWARTZ S H. Are there universal aspects in the structure and contents of human values?[J]. Journal of social issues, 1994, 50(4): 19−45.

[144] LYONS S T, DUXBURY L, HIGGINS C. An empirical assessment of generational differences in basic human values[J]. Psychological reports, 2007, 101(2): 339−352.

[145] PENEDO G, MALARTIC Q, HESSLOW D, et al. The RefinedWeb dataset for falcon LLM: outperforming curated corpora with web data, and web data only[EB/OL]. (2023−06−02)[2024−01−03]. https://arxiv.org/abs/2306.01116.

[146] CASPER S, DAVIES X, SHI C, et al. Open problems and fundamental limitations of reinforcement learning from human feedback[EB/OL]. (2023−07−28)[2024−01−03]. https://arxiv.org/abs/2307.15217.

[147] DZIRI N, MADOTTO A, ZAÏANE O, et al. Neural path hunter: reducing hallucination in dialogue systems via path grounding[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2021: 2197−2214.

[148] WANG Chaojun, SENNRICH R. On exposure bias, hallucination and domain shift in neural machine translation[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020: 3544−3552.

[149] HENDRYCKS D, GIMPEL K, NOVELLO P, et al. A baseline for detecting misclassified and out-of-distribution examples in neural networks[EB/OL]. (2016−10−07)[2024−01−03]. https://arxiv.org/abs/1610.02136v3.

[150] FEINMAN R, CURTIN R R, SHINTRE S, et al. Detecting adversarial samples from artifacts[EB/OL]. (2017−03−01)[2024−01−03]. https://arxiv.org/abs/1703.00410.

[151] DUCHI J C, GLYNN P W, NAMKOONG H. Statistics of robust optimization: a generalized empirical likelihood approach[J]. Mathematics of operations research, 2021, 46(3): 946−969.

[152] MAZEIKA M, LI B, FORSYTH D A. How to steer your adversary: targeted and efficient model stealing defenses with gradient redirection[C]//Proceedings of the

39th International Conference on Machine Learning. [S. l.]: PMLR, 2022: 15241−15254.

[153] PENG Wenjun, YI Jingwei, WU Fangzhao, et al. Are you copying my model? protecting the copyright of large language models for EaaS via backdoor watermark[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto: Association for Computational Linguistics, 2023: 7653−7668.

[154] NERGIZ M E, ATZORI M, CLIFTON C. Hiding the presence of individuals from shared databases[C]//Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data. Beijing: ACM, 2007: 665−676.

[155] BEIGI G, LIU Huan. A survey on privacy in social media[J]. IMS transactions on data science, 2020, 1(1): 1−38.

[156] LIU Kun, TERZI E. Towards identity anonymization on graphs[C]//Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. Vancouver: ACM, 2008: 93−106.

[157] MOURBY M, MACKEY E, ELLIOT M, et al. Are ‘pseudonymised’ data always personal data? Implications of the GDPR for administrative data research in the UK[J]. Computer law & security review, 2018, 34(2): 222−233.

[158] LI Bo, QI Peng, LIU Bo, et al. Trustworthy AI: from principles to practices[J]. ACM computing surveys, 2023, 55(9): 1−46.

[159] JI Shouling, MITTAL P, BEYAH R. Graph data anonymization, de-anonymization attacks, and de-anonymizability quantification: a survey[J]. IEEE communications surveys & tutorials, 2017, 19(2): 1305−1326.

[160] JI S, LI Weiqing, MITTAL P, et al. SecGraph: a uniform and open-source evaluation system for graph data anonymization and de-anonymization[C]//Proceedings of the 24th USENIX Conference on Security Symposium. [S. l.]: USENIX Association, 2015: 303−318.

[161] BEIGI G, SHU Kai, GUO Ruocheng, et al. Privacy preserving text representation learning[C]//Proceedings of the 30th ACM Conference on Hypertext and Social Media. Hof: ACM, 2019: 275−276.

[162] FEYISETAN O, BALLE B, DRAKE T, et al. Privacy- and utility-preserving textual analysis via calibrated multivariate perturbations[C]//Proceedings of the 13th International Conference on Web Search and Data Mining. Houston: ACM, 2020: 178−186.

[163] FERNANDES N, DRAS M, MCIVER A. Generalised differential privacy for text document processing[M]// Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019: 123−148.

[164] YU Da, NAIK S, BACKURS A, et al. Differentially private fine-tuning of language models[EB/OL]. (2021−10−13)[2024−01−03]. https://arxiv.org/abs/2110.06500v2.

[165] PLANT R, GIUFFRIDA V, GKATZIA D. You are what you write: preserving privacy in the era of large language models[EB/OL]. (2022−04−20)[2024−01−03]. https://arxiv.org/abs/2204.09391.

[166] ZHAO Chuan, ZHAO Shengnan, ZHAO Minghao, et al. Secure multi-party computation: theory, practice and applications[J]. Information sciences, 2019, 476: 357−372.

[167] AGRAWAL N, SHAHIN SHAMSABADI A, KUSNER M J, et al. QUOTIENT: two-party secure neural network training and prediction[C]//Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. London: ACM, 2019: 1231−1247.

[168] YANG Qiang, LIU Yang, CHENG Yong, et al. Vertical federated learning[M]//Synthesis Lectures on Artificial Intelligence and Machine Learning. Cham: Springer International Publishing, 2020: 69−81.

[169] YANG Qiang, LIU Yang, CHEN Tianjian, et al. Federated machine learning[J]. ACM transactions on intelligent systems and technology, 2019, 10(2): 1−19.

[170] HE J, XIA Mengzhou, FELLBAUM C, et al. Mabel: attenuating gender bias using textual entailment data[C]// Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2022: 9681−9702.

[171] KANEKO M, BOLLEGALA D. Debiasing pre-trained contextualized embeddings[C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Stroudsburg: Association for Computational Linguistics, 2021: 1256−1266.

[172] WEBSTER K, WANG Xuezhi, TENNEY I, et al. Measuring and reducing gendered correlations in pre-trained models[EB/OL]. (2020−10−12)[2024−01−03]. https://arxiv.org/abs/2010.06032v2.

[173] GUO Yue, YANG Yi, ABBASI A. Auto-debias: debiasing masked language models with automated biased prompts[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin: Association for Computational Linguistics, 2022: 1012−1023.

[174] ZHOU Chunting, LIU Pengfei, XU Puxin, et al. LIMA: less is more for alignment[EB/OL]. (2023−05−19)[2024−01−03]. https://arxiv.org/abs/2305.11206.

[175] GARDENT C, SHIMORINA A, NARAYAN S, et al. Creating training corpora for NLG micro-planners[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver: Association for Computational Linguistics, 2017: 179−188.

[176] PARIKH A, WANG Xuezhi, GEHRMANN S, et al. ToTTo: a controlled table-to-text generation dataset[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg: Association for Computational Linguistics, 2020: 1173−1186.

[177] RANZATO M, CHOPRA S, AULI M, et al. Sequence level training with recurrent neural networks[EB/OL]. (2015−11−20)[2024−01−03]. https://arxiv.org/abs/1511.06732.

[178] SCHULMAN J. Reinforcement learning from human feedback: progress and challenges[EB/OL]. (2023−09−27) [2024−01−03]. https://www.youtube.com/watch?v=hhiLw5Q_UFg.

[179] LI K, PATEL O, VIÉGAS F, et al. Inference-time intervention: eliciting truthful answers from a language model[EB/OL]. (2023−06−06)[2024−01−03]. https://arxiv.org/abs/2306.03341v6.

[180] SHI Weijia, HAN Xiaochuang, LEWIS M, et al. Trusting your evidence: hallucinate less with context-aware decoding[EB/OL]. (2023−03−24)[2024−01−03]. https://api.semanticscholar.org/CorpusID:258866080.

[181] KARPUKHIN V, OGUZ B, MIN S, et al. Dense passage retrieval for open-domain question answering[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2020: 6769−6781.

[182] 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. New York: Curran Associates Inc. , 2020: 9459−9474.

[183] IZACARD G, GRAVE E. Leveraging passage retrieval with generative models for open domain question answering[C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Stroudsburg: Association for Computational Linguistics, 2021: 874−880.

[184] PENG Baolin, GALLEY M, HE Pengcheng, et al. Check your facts and try again: improving large language models with external knowledge and automated feedback[EB/OL]. (2023−02−24)[2024−01−03]. https://api.semanticscholar.org/CorpusID:257205781.

[185] LUO Ziyang, XU Can, ZHAO Pu, et al. Augmented large language models with parametric knowledge guiding[EB/OL]. (2023−05−08)[2024−01−03]. https://arxiv.org/abs/2305.04757.

[186] YU W, ITER D, WANG Shuohang, et al. Generate rather than retrieve: large language models are strong context generators[C]//The Eleventh International Conference on Learning Representations. Kigali: DBLP, 2022.

[187] MADAAN A, TANDON N, GUPTA P, et al. Self-refine: iterative refinement with self-feedback[EB/OL]. (2023−03−30)[2024−01−03]. https://api.semanticscholar.org/CorpusID:257900871.

[188] MENG K, BAU D, ANDONIAN A, et al. Locating and editing factual associations in GPT[EB/OL]. (2022-02-10)[2024-01-03]. https://arxiv.org/abs/2202.05262v5.

[189] MITCHELL E, LIN C, BOSSELUT A, et al. Memory-based model editing at scale[C]//Proceedings of the 39th International Conference on Machine Learning. [S. l.]: PMLR, 2022: 15817−15831.

[190] DONG Qingxiu, DAI Damai, SONG Yifan, et al. Calibrating factual knowledge in pretrained language models[C]// Findings of the Association for Computational Linguistics: EMNLP 2022. Abu Dhabi: Association for Computational Linguistics, 2022: 5937−5947.

[191] MITCHELL E, LIN C, BOSSELUT A, et al. Fast model editing at scale[C]//International Conference on Learning Representations. Virtual Event: ICLR, 2021: 25−29.

[192] GEVA M, SCHUSTER R, BERANT J, et al. Transformer feed-forward layers are key-value memories[C]// Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2021: 5484−5495.

[193] MENG K, SHARMA A S, ANDONIAN A J, et al. Mass-editing memory in a transformer[C/OL]//The 11th International Conference on Learning Representations. (2023-02-02)[2024-01-03]. https://openreview.net/forum?id=MkbcAHIYgyS.

[194] ZHONG Zexuan, WU Zhengxuan, MANNING C D, et al. MQuAKE: assessing knowledge editing in language models via multi-hop questions[EB/OL]. (2023−05−24)[2024−01−03]. https://arxiv.org/abs/2305.14795v3.

[195] YAO Yunzhi, WANG Peng, TIAN Bozhong, et al. Editing large language models: problems, methods, and opportunities[EB/OL]. (2023−05−23)[2024−01−03]. https://arxiv.org/abs/2305.13172.

[196] HOELSCHER-OBERMAIER J, PERSSON J, KRAN E, et al. Detecting edit failures in large language models: an improved specificity benchmark[C]//Findings of the Association for Computational Linguistics: ACL 2023. Toronto: Association for Computational Linguistics,

2023: 11548−11559.

[197] ROLLER S, DINAN E, GOYAL N, et al. Recipes for building an open-domain chatbot[C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Stroudsburg: Association for Computational Linguistics, 2021: 300−325.

[198] DINAN E, HUMEAU S, CHINTAGUNTA B, et al. Build it break it fix it for dialogue safety: robustness from adversarial human attack[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong: Association for Computational Linguistics, 2019: 4537−4546.

[199] SUN Hao, XU Guangxuan, DENG Jiawen, et al. On the safety of conversational models: taxonomy, dataset, and benchmark[C]//Findings of the Association for Computational Linguistics: ACL 2022. Dublin: Association for Computational Linguistics, 2022: 3906−3923.

[200] KESKAR N, MCCANN B, VARSHNEY L, et al. CTRL: a conditional transformer language model for controllable generation[EB/OL]. (2019−09−11)[2024−01−03]. https://api.semanticscholar.org/CorpusID:202573071.

[201] LI X L, LIANG P. Prefix-tuning: optimizing continuous prompts for generation[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Stroudsburg: Association for Computational Linguistics, 2021: 4582−4597.

[202] LI X L, THICKSTUN J, GULRAJANI I, et al. Diffusion-LM improves controllable text generation[EB/OL]. (2022−05−28)[2024−01−03]. https://arxiv.org/abs/2205.14217.

[203] BAI Yuntao, KADAVATH S, KUNDU S, et al. Constitutional AI: harmlessness from AI feedback[EB/OL]. (2022−12−15)[2024−01−03]. https://arxiv.org/abs/2212.08073.

[204] THORNE J, VLACHOS A, CHRISTODOULOPOULOS C, et al. FEVER: a large-scale dataset for fact extraction and VERification[C]//Proceedings of the 2018 Conference of the North American Chapter ofthe Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans: Association for Computational Linguistics, 2018: 809−819.

[205] BERGLUND L, STICKLAND A C, BALESNI M, et al. Taken out of context: on measuring situational awareness in LLMs[EB/OL]. (2022−12−15)[2024−01−03]. https://arxiv.org/abs/2309.00667.

[206] DUAN Jiafei, YU S, TAN Hui li, et al. A survey of embodied AI: from simulators to research tasks[J]. IEEE transactions on emerging topics in computational intelligence, 2022, 6(2): 230−244.

[207] BROHAN A, BROWN N, CARBAJAL J, et al. RT-2: vision-language-action models transfer web knowledge to robotic control[EB/OL]. (2023−07−29)[2024−01−03]. https://arxiv.org/abs/2307.15818.

[208] TANNO R, PRADIER M, NORI A, et al. Repairing neural networks by leaving the right past behind[J]. Advances in neural information processing systems, 2022, 35: 13132−13145.

[209] CHO H, KIM H J, KIM J, et al. Prompt-augmented linear probing: scaling beyond the limit of few-shot in-context learners[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. [S. l.]: AAAI Press, 2023: 12709−12718.

[210] WANG Guanzhi, XIE Yuqi, JIANG Yuanfan, et al. Voyager: an open-ended embodied agent with large language models[J]. (2023−05−25)[2024−01−03]. https://arxiv.org/abs/2305.16291.

作者简介：

黄河燕，教授，兼任北京市海量语言信息处理与云计算应用工程技术研究中心主任，主要研究方向为机器翻译和自然语言处理，主持承担了国家重点研发计划项目、国家自科科学基金重点项目、国家高技术研究发展计划课题等 20 多项国家级科研攻关项目，获得国家科技进步一等奖等 10 余项国家级和省部级奖励，1997 年享受国务院政府特殊津贴，2014 年当选“全国优秀科技工作者”。E-mail：hhy63@bit.edu.cn。

李思霖，硕士，主要研究方向为信息抽取与语言模型安全性。E-mail：lisilin87@outlook.com。

郭宇航，讲师，主要研究方向为自然语言处理、信息抽取、机器翻译、机器学习、人工智能。E-mail：guoyuhang@bit.edu.cn。