

生成式大模型引发的隐私风险及治理路径

李亚玲^{1,2,3}, 蔡京京^{1,2,3}, 柏洁明^{1,2,3}

(1. 之江实验室发展战略与合作中心, 浙江 杭州, 311121;

2. 浙江省哲学社会科学试点实验室, 浙江 杭州, 311121;

3. 之江实验室智能社会治理实验室, 浙江 杭州, 311121)

摘要: 大语言模型(简称大模型)正引发人工智能技术的新一轮革命。由于大模型具有平台性、海量数据依赖性、用户交互频繁、易受到攻击等特征,其隐私泄露风险更加引人担忧。重点探讨了大模型引发的隐私泄露来源、内在机理。在梳理了国内外对大模型隐私风险的治理经验后,提出了包含政策、标准、数据、技术和生态在内的五维治理框架。最后,展望了大模型在多模态、智能体、具身智能、边缘智能等发展趋势下,可能面临的新型隐私泄露风险。

关键词: 大语言模型; 隐私; 数据泄露; 治理

中图分类号: TP39

文献标志码: A

doi: 10.11959/j.issn.2096-0271.202431

Privacy risks induced by generative large language models and governance paths

LI Yaling^{1,2,3}, CAI Jingjing^{1,2,3}, BAI Jieming^{1,2,3}

1. Development Strategy and Cooperation Center, Zhejiang Laboratory, Hangzhou 311121, China

2. Zhejiang Labs for Philosophy and Social Sciences, Hangzhou 311121, China

3. Laboratory for Intelligent Society and Governance in Zhejiang Laboratory, Hangzhou 311121, China

Abstract: Large language models(LLM) are leading to a new revolution in artificial intelligence technology. As LLM have the characteristics of platform, mass data dependence, frequent user interaction, and easy to be attacked, the risk of privacy disclosure is more worrying. The source and internal mechanism of privacy leakage caused by LLM were focused. After reviewing the domestic and foreign experience in the governance of LLM privacy risks, a five-dimensional governance framework that included policies, standards, data, technology, and ecology were proposed. Finally, this paper also looks forward to the new privacy disclosure risks that LLM may face under the development trend of multimodal, agent, embodied intelligence, edge intelligence, etc.

Key words: large language model, privacy, data leakage, government

0 引言

生成式大模型通常是指具有庞大参数规模,在海量数据上依托强大算力进行训练后,能完成不同

下游任务的机器学习模型。模型从人类提供的指令中提取和理解意图信息,并根据提取的意图来生成内容,包括文本、图像、音频、视频等。大模型以强大的学习能力、泛化能力及可扩展性打破了人工

收稿日期: 2024-04-15; 修回日期: 2024-04-28

通信作者: 蔡京京, caijj@zhejianglab.org

基金项目: 国家重点研发计划(No.2022YFC3303103); 浙江省科技厅软科学计划研究项目(No.2024C35035)

Foundation Items: National Key Research and Development Program of China(No.2022YFC3303103), Zhejiang Provincial Department of Science and Technology Soft Science Research Program(No.2024C35035)

智能技术发展的上限，加速推动生产力变革。尽管大模型因其强大的内容生成能力和生成效率引起了广泛关注，但其大规模应用仍存在一些潜在风险，如隐私泄露。大模型本身具有平台性、海量数据依赖性、用户交互频繁、易受到攻击等特征，与传统人工智能技术相比，大模型因其训练数据涉及的范围广、总量大，其引发的隐私泄露风险更加突出和引人担忧^[1]。近年来，已发生多起因大模型隐私泄露而产生的纠纷。2023年7月，美国联邦贸易委员会（Federal Trade Commission, FTC）对OpenAI发起调查，以检查ChatGPT相关服务是否侵犯用户的隐私权。2023年12月，《纽约时报》向OpenAI及微软提起侵犯版权诉讼，指控其未经许可使用其数百万篇文章以训练大模型。

上述现状表明，如何促进生成式大模型的发展与应用，同时防范其隐私泄露风险，推动技术沿着有利于人类福祉的方向和谐发展，已成为亟待思考和解决的重要命题。目前，围绕生成式大模型的隐私泄露风险已存在一些研究。Plant等^[2]表明，大模型越复杂，预训练中使用的数据量越大，其数据泄露风险越高。Emmery等^[3]表明，生成式大模型可能会在与用户的交互过程中泄露训练数据。Chen等^[4]以大型语言模型为例表明，攻击者可以准确地从模型中恢复训练数据。Carlini等^[5]在针对扩散模型的研究中，从扩散模型中成功提取了1 000多个训练样本，包括个人照片、公司徽标等。褚乐阳等^[6]以教育领域为例，探讨了大模型应用对数据隐私保护带来的挑战，主要源于大模型对数据集的依赖、自然语言交流方式及交付方式所增加的隐私泄露风险。许志伟等^[7]从评测的角度，研究了大模型安全性和隐私性的评测挑战及应对之道，为量化评估大模型的隐私风险提供了可行的路径。总体而言，上述研究多集中于技术角度的分析，缺乏技术治理视角的研究，本文将生成式大模型的隐私泄露风险与治理路径的研究结合起来，以提出更加具有针对性的治理框架。

1 大模型引发的隐私风险来源

从隐私泄露机理的角度来看，生成式大模型引发的隐私风险可以从数据收集、模型训练、模型应用3个层次来分析。来源于数据收集环节的隐私泄露风险通常与大模型对海量训练数据的需求有关；来源于模型训练环节的隐私泄露风险与大模型对敏

感数据的记忆，以及在联合训练场景下的数据流动有关；来源于模型应用环节的隐私泄露风险与模型对交互数据的过度收集、模型本身的强推理能力以及多种外部恶意攻击方式有关。

1.1 来源于数据收集环节的隐私泄露风险

生成式大模型需要在大规模数据上进行训练。以Meta最新发布的开源模型Llama 3为例，其训练数据集超过15T token。这些训练数据通常包含维基百科、书籍、期刊、Reddit网站数据、Common Crawl语料库和其他数据集。在这些数据中，可能包含用户个人信息，如姓名、年龄、性别、职业、兴趣爱好、社交关系等。如果隐私数据混入训练集，可能为后续的隐私泄露埋下隐患。

1.2 来源于模型训练环节的隐私泄露风险

在模型训练环节，大模型对训练数据的记忆，以及联合训练过程中的数据共享和流动，都有可能引发隐私泄露风险。

（1）模型的记忆问题

生成式大模型通过在海量训练数据上学习语言知识，形成强大的文本生成能力。文献[8]已经证明，大模型会记忆部分训练数据。当输入适当的提示时，这些记忆的数据可以被逐字提取。记忆与模型大小、数据重复呈线性对数关系。随着参数规模的增长，模型对训练数据的记忆问题会变得更加严重。当模型规模远大于训练数据规模时，模型就有足够的表达能力去逐字“背诵”训练样本，而不是去学习样本背后的一般化规律。文献[9]指出，15亿参数量的GPT-2模型对训练数据集的记忆比例大约为0.00000015%；而60亿参数量的GPT-J模型至少记忆了其训练数据集的1%。记忆还与训练数据中的重复序列有关，即重复出现次数越多的字符串越容易被记住。攻击者可以利用大模型对训练数据的记忆来提取训练数据（参见第1.4节）。

（2）模型联合训练时的数据流动

在医疗、金融、政务等敏感应用场景中，往往涉及大模型的联合训练和开发。在这个过程中，存在用户数据、模型参数在多方之间频繁流动的情况。如果缺乏安全的传输通道和隐私保护机制，这一过程可能会导致数据泄露、被窃取。

1.3 来源于模型应用环节的隐私泄露风险

在模型应用环节，由于用户与模型间存在频繁的用户交互，可能会发生主动或被动的隐私泄露。此外，一旦模型进入部署应用环节，将不可避免地

被外在的安全攻击所影响,部分攻击会引发训练数据泄露,进而导致隐私泄露。

(1) 模型对用户交互数据的过度收集

当用户在不知情的情况下与大模型共享个人或敏感数据时,可能会引发隐私泄露风险。例如,部分用户在与大模型交互过程中可能输入涉及工作文档、个人信息等隐私数据。尽管 ChatGPT 等模型声称不会收集这些信息,但仍然存在被攻击或窃取的风险。

(2) 模型推理侵犯隐私

随着大模型推理能力的增强,大模型能够从推理时提供给它们的大量非结构化文本集合(如公共论坛或社交网络帖子)中自动推断出广泛的个人信息。这可能会使大模型造成的隐私风险进一步扩大,而不再限于记忆问题。在一项最近的研究中^[10],研究者通过 Reddit 个人资料构建了一个文本数据集,然后让大模型“玩推理游戏”推断其中的个人信息,实现了85%的 Top-1 准确率和95.8%的 Top-3 准确率,成本仅仅是人类社会工程的1%。

1.4 外部攻击引发隐私泄露

(1) 成员推断攻击

成员推断攻击的目的是推断某个样本是否属于模型的训练数据集。成员推断攻击的核心思想是利用模型对训练样本和测试样本的处理差异性。一般而言,模型在训练样本上的表现通常要优于测试样本,如更高的置信度、更小的损失值等。攻击者可以精心设计一些探针样本,观察模型对这些样本的预测结果,从而判断它们是否在训练集中出现过。一旦攻击者发现探针样本与某个用户的隐私数据高度吻合,就可以推断该隐私数据为训练样本,即被模型记忆。

大模型之所以容易受到成员推断攻击,主要有以下原因:其一,大模型的训练数据通常包含大量长文本序列,如文章、对话、评论等。这些长文本序列数据携带了更丰富的上下文信息,使某些隐私片段更容易被关联和还原;其二,大模型往往具有强大的记忆能力,能够在海量知识的长期存储与快速检索间实现平衡。这也意味着那些高度敏感和特异的隐私信息更容易被模型牢牢记住,在推理阶段被触发;其三,大模型常以开放应用程序接口(application program interface, API)、模型共享等形式对外开放,供第三方进行各种创新应用。这虽然极大地赋能了全社会的智能化进程,但也为隐私

数据的被动泄露埋下隐患。攻击者可以频繁地查询 API,针对性地设计探针,从而增大成员推断的成功率。

(2) 数据提取攻击

数据提取攻击旨在无差别地提取训练数据,而不是从训练数据中重建特定的代表性样本。由于大模型对训练数据的记忆效应,数据提取攻击很容易导致大模型的训练数据隐私泄露。文献[8]已经证明,可以对 GPT-2 模型执行训练数据提取攻击,通过查询模型恢复数百个训练示例,包含个人身份信息(姓名、电话号码和邮件地址)、互联网中继聊天对话数据、代码和128位通用唯一识别码(universally unique identifier, UUID)等信息。具体攻击原理是通过无条件地从 GPT-2 采样来生成大量数据;然后,计算 GPT-2 生成序列的困惑度与阴影模型之间的似然比,删除不太可能包含训练数据的生成样本。对于经过“对齐”的大模型,常规的数据提取方法往往难以奏效。Nasr 等^[11]提出了一种发散数据提取方法,可以有效提取主流模型(包含对齐的闭源模型)的训练数据,如 LLaMA-65B、GPT-Neo、GPT-2、GPT-3.5-instruct 等。该攻击会导致大模型偏离预设的聊天机器人风格,使其表现得像一个基础语言模型,以比正常行为高150倍的速度快速输出训练数据。

(3) 模型窃取

与前述几种攻击方式不同,模型窃取是基于模型的输入-输出对,推测甚至复制模型内部的参数或结构,进而构建目标模型的副本。当前,大多数商业大模型开放了 API 调用的访问形式,这些 API 允许用户使用模型,但无法获知其参数。文献[12]表明通过 API 提供的大模型容易被攻击者发送大量查询请求而实现模型窃取。在谷歌的一项最新研究中^[13],研究人员通过不到2 000次的 API 查询,就成功还原了 OpenAI GPT 系列两个基础模型 Ada 和 Babbage 的整个投影矩阵,并确定了两个模型的隐藏维度(分别为1 024和2 048)。这种攻击方法的成本极低:攻击 Ada 和 Babbage 模型的成本不足20美元,攻击 GPT-3.5 的成本约200美元。除了窃取模型本身的知识产权外,模型窃取还可能会导致训练数据中的敏感信息泄露。

2 大模型的隐私风险治理现状

如何最大程度发挥生成式大模型技术效能的同

时，减少新兴技术对社会发展造成的负面冲击，已成为全球关注的重点议题。世界主要国家都在积极推动大模型创新发展应用，同时大力加强对以隐私泄露风险为代表的各类风险的前瞻研究和主动预防，力图在新一轮人工智能发展浪潮中占得先机、赢得主动。

美国在2023年10月发布白宫首个生成式大模型监管规定《关于安全、可靠、值得信赖地开发和使用人工智能的行政命令》（以下简称《命令》），以确保人工智能系统安全、可靠。《命令》中专门针对隐私保护列出了小节，指出联邦政府将确保数据的收集、使用和保留合法安全，以降低隐私泄露风险。2024年2月，美国宣布成立人工智能安全研究所联盟，该联盟包含200多家企业、高校、政府机构等单位。该联盟的主要任务之一是制定红队攻击、安全监测评估等方面的指导方针，提升大模型技术的安全性，进而减少隐私泄露等风险发生。欧盟早在2016年就通过了《通用数据保护条例》，可直接适用于包含大模型在内的人工智能技术。2024年3月，欧盟正式投票通过并批准《人工智能法案》。该法案强化了对个人隐私数据的保护要求，强调开发者必须遵循数据最小化原则，只收集完成任务所必需的数据，并采取措施保护数据的机密性、完整性和可用性。中国于2023年7月颁布《生成式人工智能服务管理暂行办法》，强调服务提供者在涉及个人信息时，应当依法承担个人信息处理者责任，履行个人信息保护义务。

除了政府机构外，国外头部科技公司针对大模型的隐私风险治理也已开始布局。微软于2023年8月公布了全新的AI服务条款，明确强调将不会保留用户对话内容，也不会将对话数据用于训练AI模型。为了应对欧盟对数据隐私的监管风险，OpenAI在都柏林新设法律实体，作为欧洲用户数据的控制者，并更新了ChatGPT产品的个人隐私条款，以符合欧盟的法律条款。

3 大模型的隐私风险治理框架

大模型的隐私风险治理是一项系统工程，需要从政策、标准、数据、技术和生态等维度多管齐下。通过完善相关政策和标准，加强数据治理，引入安全保障技术，汇聚多方力量，构建多方合作的治理生态，从而共同推动技术向善，发展负责任的人工智能。

3.1 政策

一是建立促进大模型发展的专项法律法规，形成大模型隐私风险治理的顶层纲领。二是厘清训练数据的版权争议，补充关于大模型训练数据合理使用的法律规定。据不完全统计，2022年11月至2023年10月，仅美国加州北区法院便已经受理了10起版权人起诉Stability AI、OpenAI等大模型企业未经授权，利用版权作品进行模型训练的案件。当前“先授权后使用”的版权法律体系已难以适用大模型时代的数据使用方式。为解决上述矛盾，可尝试在《著作权法》“合理使用”情形中增加“文本与数据挖掘”例外条款，正面回应大模型训练数据使用问题。三是健全大模型应用场景下的个人信息保护实施细则。可选取传媒、教育、医疗等大模型成熟应用的领域，开展试点探索，摸清关键环节的隐私风险问题，完善大模型隐私信息的泄露监测和风险评估细则。

3.2 标准

标准层面，我国已出台针对算法的安全评估国家标准，如《信息安全技术机器学习算法安全评估规范》《大规模预训练模型技术和应用评估方法》等，但针对大模型的安全评估标准尚未制定。未来，可进一步完善大模型全生命周期的隐私保护标准，涵盖数据采集、清洗、存储、使用、共享、销毁等环节；规范身份认证、访问控制、数据加密、日志审计等数据安全标准；制定大模型API的身份认证、权限管理、流量控制等安全标准。基于上述标准，形成大模型的安全标准规范，引导技术安全可靠发展。

3.3 数据

海量、异构、非结构化的互联网数据虽为大模型提供了丰富的知识来源，但也使隐私数据泄露、滥用的风险倍增。加强对大模型训练数据的隐私治理至关重要，可采取的措施如下。

（1）数据去重

语言模型的训练数据集往往存在着显著的序列级重复。这种重复会导致大模型的“记忆”问题，并容易受到隐私攻击。文献[14]提出，大模型生成训练序列的频率与序列的重复出现次数呈超线性相关。例如，在训练数据中出现10次的序列平均生成频率比仅出现一次的序列高1 000倍。Lee等^[15]进一步表明，训练数据去重可以显著提高训练效率，同时并不会降低模型性能。

(2) 数据过滤

数据过滤旨在在模型训练之前删除训练数据中的敏感信息,如个人身份信息。对于大型文本数据集,通常基于模式解析器或分类算法等文本自动识别方法选中并删除特定的敏感序列^[16]。但这种方法存在一些挑战。一方面,界定哪些数据属于隐私信息比较困难。有些隐私信息依赖于上下文,并且有多种形式和格式,难以识别。另一方面,在敏感信息难以完全过滤的情况下,随着模型推理能力的提升,哪怕有少量敏感信息遗漏,也可以通过模型推理恢复,从而导致隐私泄露。因此,数据过滤方法一般需要与其他隐私保护方法相结合。

(3) 数据水印

在训练数据中植入隐蔽的水印信息,一旦发生隐私数据泄露事件,可根据水印快速溯源,并采取应对措施。这种主动防御策略可以遏制隐私数据的二次传播。Sablayrolles等^[17]最早引入了放射性数据的概念,通过给训练集的图像添加水印,在后续对其进行训练的分类器中留下可检测的痕迹,从而加强数据的可追溯性。Sander等^[18]进一步将数据水印技术用于检测大模型A的生成文本是否被用于大模型B的微调。结果表明,即使只有5%的训练文本带有水印,该方法仍可以高置信度(p 值 $<10^{-5}$)地检测到加水印的训练数据留下的痕迹。

3.4 技术

从技术层面防范大模型的隐私风险,主要有以下措施。

(1) 联邦学习

在医疗和金融等特殊领域,数据隐私保护等级较高,训练数据往往难以公开共享。联邦学习(federated learning, FL)提供了一种分布式机器学习范式,可实现在分散隐私数据上的协作训练,从而解决上述问题^[19]。与使用集中式训练方法的传统机器学习相比,FL通过在没有本地原始数据的情况下循环和处理中间加密数据来实现联合多方学习训练。在训练过程中,无须将任何私人数据上传到服务器,从而有效降低了与传统机器学习源的数据聚合相关的隐私风险。Ye等^[20]提出的OpenFedLLM框架,集成了联邦指令调优、联邦值对齐等关键功能。研究发现,在一些需要特定领域专家知识的特定领域,如金融,FL微调的LLaMA2-7B在相应数据集上的表现甚至可以超过GPT-4。

(2) 安全多方计算

安全多方计算(secure multi-party computation, SMC)是密码学中一种保护隐私的分布式计算技术。SMC技术可以在推理过程中保护数据和模型权重的隐私,从而在不泄露原始数据内容的情况下捕捉数据使用的价值。由于Transformer架构中存在大量非线性操作,直接应用SMC技术会导致模型性能的大幅下降。已有许多研究致力于Transformer架构的安全推理研究。主要有两种技术路线,一种技术路线是基于Transformer算子替换的优化,例如THE-X框架^[21]、MPCFormer框架^[22]和MPCViT框架^[23],主要思路是将Transformer架构中的复杂计算函数(如GeLU函数、Softmax函数)替换为简单函数,以简化安全计算协议,减少计算开销。但这类技术路线一般需要对简化后的模型进行再次训练以提升模型性能。另一种技术路线是基于安全计算协议的优化,例如Iron框架^[24]、PUMA框架^[25]和SIGMA框架^[26]。PUMA框架通过对GeLU函数和Softmax函数的高精度近似,在保持模型性能的同时显著降低了安全推理的成本,是首个实现LLaMA-7B级别模型安全推理的方案。

(3) 基于硬件的可信执行环境

基于硬件的可信执行环境(trusted execution environment, TEE)技术通过在CPU中封装一个安全隔离区,确保敏感数据和代码只在这个可信区域内执行,即使系统被攻陷,TEE内部数据也难以被获取。TEE基于硬件安全CPU,能够实现与内存隔离的安全计算,在保持计算效率的同时实现隐私保护计算。主流的TEE实现方案包括Intel SGX、AMD SEV、英特尔TDX和TrustZone。Huang等^[27]将联邦学习与可信执行环境结合,提出了一种基于模型切片的安全分布式大模型训练框架,在确保安全性的同时实现了几乎无损的模型准确性。

(4) 同态加密

同态加密(homomorphic encryption, HE)首先由模型所有者提供训练好的深度学习模型;其次,数据所有者/模型使用者输入加密的敏感数据,发送给模型所有者进行推理;然后由模型所有者处理加密数据,将加密的输出结果返回给数据所有者,数据所有者对其解密并获得最终的分类结果。在整个过程中,数据在加密状态下参与计算,实现了模型所有者和数据所有者的知识解耦,第三方难以从数据和计算过程中窥探隐私。现代同态加密方案(如近似同态加密算法)仅支持多项式函数表

示。因此，基于Transformer架构的大模型在应用同态加密技术时需要将非线性激活函数调整为等效的多项式形式。Zimerman等^[28]首次提出了自注意力和归一化层的多项式替代方案，拓展了基于HE的Transformer架构模型的应用潜力。

（5）差分隐私

差分隐私（differential privacy, DP）通过引入随机噪声，让模型对单个训练样本的依赖降到最低，由此防止隐私数据被还原。差异隐私可以帮助减少大模型在训练过程中记忆敏感信息的机会，但也会由于引入噪声而降低模型的性能，导致可用性和准确性下降。因此需要在隐私保护和模型功能之间寻求均衡。Shi等^[29]提出了一种新型框架JFT，为基于Transformer架构的大模型提供选择性差分隐私保护方案。结果表明，JFT生成的模型具有较强的性能，同时对金丝雀插入攻击¹保持鲁棒性。

3.5 生态

大模型的隐私风险治理是一个全生命周期的系统工程。需要推动社会共治，形成共治共享的治理生态。在隐私教育上，需要加强面向开发者的大模型隐私保护教育，提高其风险意识和责任意识，同时加强面向用户的数据权益教育，提高用户的隐私保护意识和能力。在多方参与上，鼓励多元利益相关方参与大模型的隐私风险治理，包括政府、企业、学界、公众等。搭建利益相关方对话平台，协调不同主体间的利益诉求，形成社会共治合力。在公众监督上，建立大模型“隐私众测”机制，发挥公众监督作用，及时发现隐私漏洞。鼓励大模型提供方定期发布大模型的隐私风险治理报告，接受社会各界监督。此外，大模型的隐私风险治理还需注重国际合作。隐私泄露是全球共同面临的挑战，需要加强跨国、跨区域的政策对话和务实合作。

4 挑战与未来方向

当前，大模型技术迅速迭代，多模态、AI智能体、具身智能、边缘智能成为未来大模型技术发展的重要方向。然而，这些融合系统可能会由于引入新的模态、设备和交互过程，增加大模型的隐私泄露风险。

（1）多模态大模型的信息泄露

多模态大模型通过在大规模文本-图像对、文本-视频对上进行训练，学习不同模态间的关联。其隐私泄露主要源于模态间信息的绑定、转移、放

大和关联。跨模态学习一方面赋予了模型更强大的生成能力，另一方面也加剧了隐私在不同模态间传播的风险。当某个模态的训练数据发生隐私泄露时，另一个模态可能会放大这种泄露。攻击者可以利用多个模态的信息交叉印证，获得更全面、更准确的隐私画像。例如文本到图像生成模型中，可能利用文本描述生成对应人物的面部图像，导致面部隐私泄露。

（2）大模型+智能体

大模型与AI智能体的融合正成为新的发展方向。然而，这种融合也可能扩大隐私泄露的风险。传统的大模型主要将隐私信息隐式编码在语言模型参数中。而融合后的智能体可能将隐私信息作为知识存储、推理和决策，扩大隐私影响范围。例如，在执行“医疗诊断”任务时，智能体可能将用户透露的“近期头晕乏力”等隐私症状，作为重要知识点串联问诊逻辑，并在后续多轮会话中反复提及，加剧隐私扩散。此外，大模型与智能体融合后，人机交互和任务执行能力会大大增强，可能会通过自主学习进化出意料之外的隐私窃取能力。

（3）大模型+具身智能

大模型与机器人的融合，也称为具身智能，有望大幅提升机器人在复杂环境中的自主性、适应性、交互性，助力其在工业、服务业等领域的广泛应用。但其引发的隐私风险也不容小觑。一方面，具身智能赋予大模型视觉、听觉、触觉等多模态感知能力，可以更全面地洞察用户的行为、情绪、习惯等隐私信息。例如家用服务机器人可以通过视觉系统捕捉用户的日常活动，通过语音系统捕捉用户的对话内容，通过触觉系统捕捉用户的身体接触等。一旦这些具身感知数据被上传至云端进行大模型学习，用户隐私将更加透明。另一方面，具身智能赋予了大模型更持久的任务记忆能力。传统的大模型主要依赖短期的上下文记忆来维持对话，很少有长期的任务存储。而与机器人系统的结合，尤其是采用强化学习等目标导向的学习范式，使其获得了跨场景、跨任务的长期记忆能力，可以在执行任务的过程中持续积累和利用隐私数据。

（4）大模型+边缘智能

1 金丝雀插入攻击（canary insertion attack）：在训练数据中插入特定的字段，如论文中插入“My ID is 341 752”，并计算该字段的暴露度，暴露度越低，模型越安全。

随着模型轻量化和边缘算力提升,大模型未来在智能手机、个人计算机等边缘端的渗透将更快。谷歌最新发布的开源大模型 Gemma 支持在个人计算机上运行。微软将在下一代个人计算机上增设实体 Copilot 键,用于快速调用大模型助手。这种边缘智能范式一方面可以极大提升本地智能服务的时效性、个性化和隐私性,减少数据上云带来的隐私泄露风险;另一方面,由于边缘设备的计算、存储、安全防护等能力有限,其脆弱性也可能引发新的隐私泄露隐患。大模型的海量参数给边缘设备的存储和安全管理带来沉重负担。一旦边缘设备失陷,其掌握的隐私数据将完全泄露。

针对上述大模型未来发展中动态变化的隐私风险,除了基于政策、标准、数据、技术和生态的五维治理框架外,还需要持续监控大模型发展中的隐私风险变化特征,完善风险监测和预警机制,促进多方协作和信息共享,以确保及时识别和减轻隐私风险。在人机共生、虚实共融的未来中,既要在大模型的应用潜力保持乐观,又要对其风险保持警惕。需适度超前研究未来技术发展趋势下可能引入的隐私风险,以提早防范,引导技术向善和造福人类。

参考文献:

- [1] 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.
- [2] PLANT R, GIUFFRIDA V, GKATZIA D. You are what you write: preserving privacy in the era of large language models[EB]. arXiv preprint, 2022, arXiv: 2204.09391.
- [3] EMMERY C, KÁDÁR Á, CHRUPALA G. Adversarial stylometry in the wild: transferable lexical substitution attacks on author profiling[C]// *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Stroudsburg: Association for Computational Linguistics, 2021.
- [4] CHEN S, KAHLA M, JIA R X, et al. Knowledge-enriched distributional model inversion attacks[C]// *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Piscataway: IEEE Press, 2021: 16158-16167.
- [5] CARLINI N, HAYES J, NASR M, et al. Extracting training data from diffusion models[EB]. arXiv preprint, 2023, arXiv: 2301.13188.
- [6] 褚乐阳, 潘香霖, 陈向东. AI大模型在教育应用中的伦理风险与应对[J]. *苏州大学学报(教育科学版)*, 2024, 12(1): 87-96.
- [7] CHU L Y, PAN X L, CHEN X D. Ethic risks and responses in the application of AI large language models in education[J]. *Journal of Soochow University(Educational Science Edition)*, 2024, 12(1): 87-96.
- [8] 许志伟, 李海龙, 李博, 等. AIGC大模型测评综述: 使能技术、安全隐患和应对[J]. *计算机科学与探索*, 2024, 18(9): 2293-2325.
- [9] XU Z W, LI H L, LI B, et al. A survey of AIGC model evaluation: enabling technologies, vulnerabilities and mitigation[J]. *Journal of Frontiers of Computer Science and Technology*, 2024, 18(9): 2293-2325.
- [10] CARLINI N, TRAMER F, WALLACE E, et al. Extracting training data from large language models[EB]. arXiv preprint, 2020, arXiv: 2012.07805.
- [11] CARLINI N, IPPOLITO D, JAGIELSKI M, et al. Quantifying memorization across neural language models[EB]. arXiv preprint, 2022, arXiv: 2202.07646.
- [12] STAAB R, VERO M, BALUNOVIĆ M, et al. Beyond memorization: violating privacy via inference with large language models[EB]. arXiv preprint, 2023, arXiv: 2310.07298.
- [13] NASR M, CARLINI N, HAYASE J, et al. Scalable extraction of training data from (production) language models[EB]. arXiv preprint, 2023, arXiv: 2311.17035.
- [14] KRISHNA K, TOMAR G S, PARIKH A P, et al. Thieves on sesame street! model extraction of bert-based apis[EB]. arXiv preprint, 2019, arXiv: 1910.12366.
- [15] CARLINI N, PALEKA D, DVIJOTHAM K D, et al. Stealing part of a production language model[EB]. arXiv preprint, 2024, arXiv: 2403.06634.
- [16] KANDPAL N, WALLACE E, RAFFEL C. Deduplicating training data mitigates privacy risks in language models[C]// *Proceedings of the 39th International Conference on Machine Learning*. New York: PMLR, 2022: 10697-10707.
- [17] LEE K, IPPOLITO D, NYSTROM A, et al. Deduplicating training data makes language models better[EB]. arXiv preprint, 2021, arXiv: 2107.06499.
- [18] WANG Z, ZHONG W, WANG Y, et al. Data management for large language models: a survey[EB]. arXiv preprint, 2023, arXiv: 2312.01700.
- [19] SABLAYROLLES A, DOUZE M, SCHMID C, et al. Radioactive data: tracing through training[C]// *Proceedings of the 37th International Conference on Machine Learning*. New York: PMLR, 2020: 8326-8335.
- [20] SANDER T, FERNANDEZ P, DURMUS A, et al. Watermarking makes language models radioactive[EB]. arXiv preprint, 2024, arXiv: 2402.14904.
- [21] 王飞跃, 王艳芬, 陈慧竹, 等. 联邦生态: 从联邦数据到联邦智能[J]. *智能科学与技术学报*, 2020, 2(4): 305-311.
- [22] WANG F Y, WANG Y F, CHEN Y Z, et al. Federated ecology: from federated data to federated intelligence[J]. *Chinese Journal of Intelligent Science and Technology*, 2020, 2(4): 305-311.
- [23] YE R, WANG W H, CHAI J Y, et al. OpenFedLLM: training large language models on decentralized private data via federated learning[EB]. arXiv preprint, 2024, arXiv: 2402.06954.
- [24] CHEN T, BAO H, HUANG S, et al. THE-X: privacy-preserving transformer inference with homomorphic encryption[EB]. arXiv preprint, 2022, arXiv: 2206.00216.
- [25] LI D, SHAO R, WANG H, et al. MPCFormer: fast, performant and private Transformer inference with MPC[EB]. arXiv preprint, 2022, arXiv: 2211.01452.
- [26] ZENG W X, LI M, XIONG W J, et al. MPCViT: searching for accurate and efficient MPC-friendly vision transformer with heterogeneous attention[C]// *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. Piscataway: IEEE Press, 2023: 5029-5040.
- [27] HAO M, LI H, CHEN H, et al. Iron: private inference on transformers[J]. *Advances in Neural Information Processing Systems*, 2022, 35: 15718-15731.

- [25] DONG Y, LU W J, ZHENG Y C, et al. PUMA: secure inference of LLaMA-7B in five minutes[EB]. arXiv preprint, 2023, arXiv: 2307.12533.
- [26] GUPTA K, JAWALKAR N, MUKHERJEE A, et al. SIGMA: secure GPT inference with function secret sharing[EB]. Cryptology ePrint Archive, 2023, cryptoeprint: 2023/1269.
- [27] HUANG W, WANG Y, CHENG A, et al. A fast, performant, secure distributed training framework for large language model[EB]. arXiv preprint, 2024, arXiv: 2401.09796.
- [28] ZIMMERMAN I, BARUCH M, DRUCKER N, et al. Converting transformers to polynomial form for secure inference over homomorphic encryption[EB]. arXiv preprint, 2023, arXiv: 2311.08610.
- [29] SHI W, SHEA R, CHEN S, et al. Just fine-tune twice: Selective differential privacy for large language models[EB]. arXiv preprint, 2022, arXiv: 2204.07667.

[作者简介]



李亚玲（1991-），女，博士，之江实验室发展战略与合作中心助理研究员，主要研究方向为人工智能技术的风险治理等。



蔡京京（1994-），女，之江实验室发展战略与合作中心算法工程师，主要研究方向为推荐算法、自然语言处理算法和大模型训练算法等。



柏洁明（1990-），男，之江实验室发展战略与合作中心算法工程师，主要研究方向为大模型应用及数据挖掘。