面向在线生成式人工智能服务的隐私保护方法

齐 涛[1],王慧丽[2],杨珮茹[2],王文丹[1],谭支鹏[3],黄永峰[2],王尚广[1],徐红艳[4*],罗传文[4]

(1. 北京邮电大学计算机学院网络与交换技术全国重点实验室,北京 100876;2. 清华大学电子工程系,北京 100084;
3. 华中科技大学武汉光电国家研究中心,湖北武汉 430074;4. 北京林业大学信息学院,北京 100083)

摘 要: 近年来,在线人工智能系统在众多领域展现出强大的推理能力,对社会产生了广泛的影响。在使用此类模型服务时,用户通常需将相关查询数据上传至云端平台以提供明确的任务指令。然而,这些查询数据可能包含隐私敏感或者机密信息,直接与云端平台共享会存在隐私泄露风险。此外,人工智能平台通常也会收集并利用用户数据进一步训练模型,可能导致用户的私有信息被生成式大模型记忆,并在后续公共服务中被生成并传播,从而加剧隐私泄露的可能性。现有生成式人工智能应用的隐私保护机制普遍依赖于针对提示词的脱敏技术,其安全性高度依赖敏感信息识别的准确性,通常需依赖大量标注数据进行隐私识别模型训练,不仅在实施成本上存在挑战,在训练过程中还极有可能引入新的隐私漏洞。为应对这一问题,本文提出一种新型隐私保护协同学习框架PrivateAI,该框架的核心思想是在严格保障隐私安全的前提下,充分利用分散在不同终端设备中的敏感数据,以训练本地隐私识别模型。同时,PrivateAI通过提取云端大模型推理过程中隐含的知识,并将其压缩为轻量级知识蒸馏数据集,实现对本地模型的高效性能增强。此外,针对标注数据和大模型蒸馏数据的异构性挑战,本框架引入了异构知识融合机制,用于对齐并整合来自基础模型与分布式标注数据的多源知识,从而显著提升隐私识别模型的泛化能力与隐私风险预警性能。为验证PrivateAI的实际效果,本文在两个真实医疗数据集上进行了系统评估。该框架能够在满足隐私约束的前提下,有效训练隐私识别模型,并对潜在隐私风险进行预警。在两个公开医疗数据集上的实验结果表明,PrivateAI训练得到的模型可最高提升53.7个百分点的隐私保护成功率。上述验证展现出PrivateAI在缓解隐私泄露风险方面的潜力,可作为在线智能应用中预防隐私泄露的有效工具。

关键词: 隐私保护;协同学习;在线人工智能服务;差分隐私;联邦学习

基金项目: 国家自然科学基金(No.62425203,No.62502044)

中图分类号: TP18;TP393.08;TP311.13 **文献标识码:** A **文章编号:** 0372-2112(2026)01-0050-18

电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250793

Building Privacy Shield in Online Generative AI Services

QI Tao[1], WANG Huili[2], YANG Peiru[2], WANG Wendan[1], TAN Zhipeng[3], HUANG Yongfeng[2], WANG Shangguang[1], XU Hongyan[4*], LUO Chuanwen[4]

(1. *State Key Laboratory of Networking and Switching Technology*, *School of Computer Science*, *Beijing University of Posts and Telecommunications*, *Beijing* 100876, *China*; 2. *Department of Electronic Engineering*, *Tsinghua University*, *Beijing* 100084, *China*;
3. *Wuhan National Optoelectronics Research Center*, *Huazhong University of Science and Technology*, *Wuhan*, *Hubei* 430074, *China*;
4. *School of Information Science*, *Beijing Forestry University*, *Beijing* 100083, *China*)

Abstract: In recent years, state-of-the-art online artificial intelligence systems demonstrate remarkable capabilities in various fields, exerting broad social impacts. In order to access these model services, users are typically required to upload their personal data to the cloud platform. However, these queries may contain sensitive or confidential information, and directly sharing them with cloud platforms introduces potential privacy leakage risks. Moreover, platforms may exploit user data for further model training, causing private information to be memorized by the model and later regenerated in public services, thereby aggravating the risk of privacy breaches. Existing privacy-preserving mechanisms in generative AI applications predominantly rely on prompt sanitization techniques, whose security critically depends on the accuracy of sensitive information identification. These approaches usually require large amounts of annotated data for model training, which not only raises implementation costs but may also introduce new privacy vulnerabilities in specific scenarios. To address this issue, this paper proposes a novel privacy-preserving collaborative learning framework named PrivateAI. The core idea of this framework is to fully exploit sensitive data distributed across different devices to train local privacy identification models, while strictly ensuring data privacy. Meanwhile, PrivateAI extracts the implicit knowledge embedded in the large foun-

收稿日期:2025-09-10;录用日期:2026-01-07;责任编辑:覃怀银

*通信作者:徐红艳

dation models and compresses it into a lightweight distilled dataset, thereby achieving effective privacy detection performance enhancement of local models. In addition, to tackle the heterogeneity challenge between the knowledge extracted from labeled data and foundation models, the framework introduces a heterogeneous knowledge fusion mechanism that aligns and integrates multi-source knowledge from both the foundational models and distributed labeled datasets. We evaluate PrivateAI on two datasets, and the results demonstrate that models learned by PrivateAI can maximally improve the privacy protection success rate by 53.7 percentage points. PrivateAI holds significant potential in mitigating privacy breaches, acting as a sentinel against severe privacy leakage incidents within online AI applications.

Keywords: privacy protection; collaborative learning; online artificial intelligence services; differential privacy; federated learning

Foundation Item(s): National Natural Science Foundation of China (No.62425203, No.62502044)

0 引言

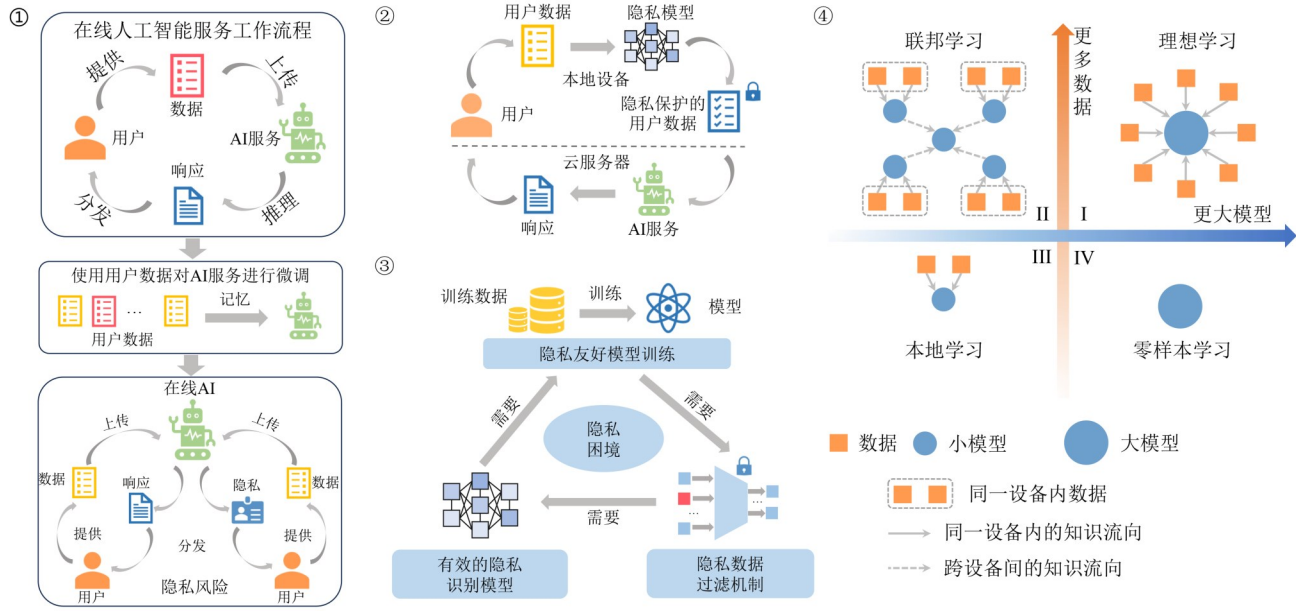
近年来,生成式人工智能技术取得了显著进步,在多个专业领域与日常应用中为用户提供了重要支持[1]。当前最先进的生成式人工智能系统大多建立在参数规模庞大的基础模型之上[2],这类模型对计算资源要求较高,导致个人用户难以在本地进行部署。因此,在线人工智能平台迅速发展,通过提供便捷可用的人工智能服务满足多样化用户需求,吸引了大量用户。以广受欢迎的人工智能聊天应用ChatGPT为例,在其推出后的两个月内月活跃用户数便突破1亿,创下了历史上用户增长最快的消费应用纪录[3]。这些在线人工智能系统不仅凭借卓越的能力受到广泛关注,也在大规模推广应用中产生了深远的社会影响。

用户提供正确、清晰的指令提示数据是生成式大模型系统能够精准执行用户任务的先决条件[4]。由于生成式大模型通常具有极高的商业价值,并且参数规模通常极大导致难以在本地设备推理计算,因此主流的在线人工智能平台通常要求用户将私人指令数据上传至远程服务器以获取相应服务。然而,在医疗问答等特定领域,指令数据可能包含敏感的用户信息,一旦发生泄漏将带来严重的隐私风险[5]。此外,人类指令数据对于改进人工智能模型具有重要价值,平台方通常会对用户提交的数据进行存储并用于模型微调[6]。这一过程可能导致模型无意中记忆用户的隐私信息,并在后续生成响应时泄露给其他用户,进一步加剧隐私泄漏问题[7](图1①)。这类风险并非仅限于理论层面,实际案例时有发生。例如,某知名商业公司的硬件工程师在使用ChatGPT时无意中泄露了关于新型硬件设备的商业机密,最终导致公司全面禁止员工使用该平台[8]。因此,构建并实施有效的隐私保护机制,对在线智能应用的持续发展具有重要意义。

实际上,一旦本地设备对外共享隐私数据,其泄露风险将显著上升,凸显了在用户边缘设备端直接部署隐私保护机制的重要性。尽管从理论上看,数据扰动等隐私增强技术能够在一定程度上保障数据安全,但这类方法往往会导致模型准确性明显下降,从而对人工智能服务的整体质量产生不利影响[9]。近期,研究者逐渐将重点转向对敏感信息的识别[10],以期在数据上传至云端之前,在本地完成敏感内容的检测与处理(图1②)。通常而言,大规模标注数据与经过预训练的大模型所蕴含的通用知识能力,是构建高效隐私识别模型的两个关键基础。在理想情况下,应基于大规模标注的隐私数据对大模型进行微调,以最优化模型的隐私识别能力。然而,个人设备通常既缺乏足够的高质量标注数据,也不具备支撑大规模模型训练所需的计算资源。此外,由于所涉及的数据通常包含敏感个人信息,若简单将分散于各设备中的标注数据集中处理以增大标注数据规模,极易引入额外的隐私泄露风险。这导致我们陷入了一种“隐私困境”,即为了实现更有效的保护,反而需要牺牲部分数据隐私(图1③)。

为解决上述隐私困境,亟需设计一种协同学习算法,能够在有效利用分散数据与大模型的同时充分保护用户隐私。一种可行的思路是通过协同训练整合来自多个设备的隐私标注数据,从而构建比单一本地训练更强大的隐私识别模型[11]。联邦学习作为一种具备一定隐私保护能力的分布式机器学习范式,允许用户在不上传原始数据的情况下参与跨设备模型训练,初步实现了上述目标(图1④:II)[12]。此外,在联邦学习中,由于跨节点共享的模型参数仍可能与原始隐私数据存在关联,因此通常需在共享前添加差分隐私噪声,以进一步提升安全性。比如,RobustDPFL方法[13]将Rényi差分隐私应用于联邦学习中,通过在本地节点模型参数上传前对其进行扰动,保护训练过程的本地数据隐私安全。DP-FTRL方法[14]通过结合正则化优化与高效噪声聚合机制,在保证差分隐私的同时缓解了隐私噪声对模型精度的破坏,实现了更优的隐私-效用平衡[15]。然而,在实际应用中,可用标注数据规模通常有限,而严格的安全机制会进一步限制训练中的知识交换效率,最终制约模型的隐私识别效

果。此外，除了标注数据，引入基础模型中已编码的通用先验知识也有助于增强隐私识别能力（图1④：IV）。然而，基础模型通常部署在云平台，难以直接通过联邦学习在用户设备之间进行共享或微调，从而导致理想训练过程与实际操作间存在显著差距（图1④：I）。因此，构建一个高效、安全的跨数据、跨模型的协同学习框架对于增强在线人工智能应用的隐私保护至关重要。



注：①在线人工智能应用中潜在的隐私泄漏风险；②基于隐私识别模型的隐私保护方法；③隐私保护困境示意图；④根据模型先验知识水平（以参数规模为指标，横轴）与训练数据量（纵轴）情况下不同隐私识别模型训练的范式示意图。

图1 在线人工智能服务中的隐私风险与保护机制

Figure 1 Privacy risks and protection methods within online AI services

为应对上述挑战，本文提出了一种新型隐私保护协同学习框架PrivateAI，旨在在充分保障用户隐私安全的同时，高效完成隐私识别模型的训练。该方法在不直接访问原始敏感数据的前提下，通过提取基础模型推理行为中编码的知识，并将其压缩至轻量级知识蒸馏数据集，实现知识的高效迁移。针对分布式设备中分散存储的隐私数据，PrivateAI仅共享基于本地数据训练得到的模型参数，而非原始数据，从而在严格保护用户隐私的基础上，充分利用分散数据的价值。此外，本文引入了一种异构知识融合机制，能够有效对齐来自基础模型与分散标注数据的知识，并将其融合到隐私识别模型中，以增强模型的隐私泄漏预警能力。与现有工作相比，PrivateAI能够高效地将大模型中编码的通用隐私识别知识迁移至本地模型中，有效缓解了本地数据不足与差分隐私干扰导致的训练低效问题，实现了训练环节隐私保护和应用环节隐私识别效果的兼得。同时，该框架在协同学习过程中协调了基于大模型蒸馏知识与分散数据训练知识之间的冲突，显著提升了模型的训练效率与整体性能，为隐私保护型协同学习提供了一种兼顾安全性与实用性的新方法。为验证本文所提出的PrivateAI方法的实际效果，本文在两个真实医疗数据集上进行了系统评估。大量实验表明，该框架在协同学习过程中能有效防止隐私泄露，隐私保护成功率可达约90%，较现有最优方法提升高达53.7个百分点。进一步的实验分析显示，该框架在实现隐私保护的同时，将来自分散数据与基础模型的异质任务知识利用效率提升至47.3%，证明了其高效整合异构任务的潜力。本文对应的实验数据、复现代码均上传至公开仓库https://github.com/taoqi98/PrivateAI。本文主要创新点如下：

（1）本文提出了一种基于本地分散标注数据与云端推理大模型协同的隐私识别模型训练范式，有效打破了大模型推理阶段的隐私困境，实现了在隐私安全约束下的高性能隐私识别模型训练，可进一步应用于模型推理环节的隐私防护。

（2）本文首次从基础大模型的推理行为中挖掘并提炼出隐私识别知识，基于代理数据集对其进行蒸馏并构建轻量化知识数据。该蒸馏知识随后被融入隐私识别模型的分布式训练过程中，有效缓解了本地标注数据稀缺与差分隐私噪声干扰等问题，从而显著提升模型的训练稳定性与识别性能。

（3）本地标注数据与基础模型蒸馏所得隐私识别

知识在表征形式与分布空间上存在显著差异(分别编码于本地模型参数与蒸馏数据集中),如何实现二者的高效融合成为关键挑战。本文提出一种基于代理数据集的表征对齐与冲突过滤机制,对齐基础模型与本地数据的知识表示,并剔除潜在冲突信息,从而实现跨源知识在目标模型中的有效融合,显著提升隐私识别模型鲁棒性。

(4)实验结果表明,在不访问原始敏感数据的分布式训练架构下,本文方法能够在严格差分隐私约束($\epsilon=1$)下实现高效的隐私识别模型训练,在隐私保护成功率与识别精度上均显著优于现有主流方案,最高性能提升可达53.7个百分点,充分验证了其在隐私增强学习场景下的有效性。

1　相关工作

1.1　基于敏感信息识别的隐私保护推理

基于敏感信息识别的隐私保护推理方法近年来受到广泛关注,其基本思路是首先识别输入文本或提示中的潜在隐私信息,然后对这些信息进行混淆、替换或扰动,最终在保证下游任务效用的同时实现隐私保护。该类方法的核心环节在于隐私信息的识别与后续的扰动机制[16]。在隐私信息识别方面,研究者提出了多种方法来实现隐私信息提取[17]。早期方法主要依赖正则表达式[18]对特定模式进行匹配,或通过关键字搜索[19]来提取与个人信息相关的内容。随后,命名实体识别[20]成为主流方向,利用BERT[21]、spaCy[22]等深度学习模型或框架识别和分类文本中的命名实体,从而实现较为精准的隐私信息定位[23]。比如,Johnson等人[24]针对临床文本的去标识化问题,提出基于BERT的双向编码器模型,通过在下游微调阶段对四个临床数据集进行训练,实现对患者姓名、日期、年龄等敏感信息的高效识别与去除。Dou等人[25]针对社交媒体自我披露场景,构建了包含19个类别4 800条标注片段的大规模自我披露数据集,并基于BERT和RoBERTa微调模型实现披露检测与抽象化,有效降低隐私风险。Mishra等人[26]提出一种混合式个人隐私信息检测与匿名化方法,结合基于规则的自然语言处理机制与定制化机器学习模型,在合成与真实文档中均实现高精度识别与泛化,提升隐私保护的可靠性与实用性。然而,为实现精准的隐私识别与脱敏,上述方法通常依赖大规模高质量的标注数据以训练模型。在实际应用场景中,标注数据规模往往有限,这一问题会严重影响上述方法的安全性与可靠性。因此,如何在缺乏大规模标注数据的条件下实现高效、稳健的隐私信息识别,仍然是一个亟待解决的问题。与现有方法相比,本文提出的方法能够在本地设备上协同利用多个分散节点的标注数据,并结合部署于云端的大语言模型所蕴含的隐私信息识别能力,以增强本地隐私识别模型的匿名化效果。该方法无需依赖大规模集中标注数据即可提升隐私识别的性能,从而显著缓解传统方法对大规模标注数据的严重依赖。

此外,随着大模型(Large Language Model,LLM)的快速发展,基于LLM的匿名化方法也逐渐兴起,它们能够在上下文中理解复杂的隐私线索,并通过自然语言改写完成匿名化任务。在整体敏感信息识别与推理的框架设计中,现有研究逐渐借助大模型的能力以进一步提升隐私保护的效果[27]。Yang等人[28]提出基于LLM的鲁棒和效用保持的文本匿名化框架RUPTA,利用LLM的序列标注能力基于提示识别文本中的敏感信息,通过隐私评估器评估隐私保护水平,提供文本反馈指导优化器迭代进行扰动操作,同时参考效用评估器的反馈保留下游任务所需信息。Staab等人[29]提出了一种基于LLM的反馈引导的对抗式匿名化方法,该方法针对传统匿名化工具只能去除显性敏感信息、难以抵御上下文隐含属性推断的问题,设计了一个循环框架,利用推断大模型尝试从文本中推测用户的敏感属性,然后使用匿名化大模型根据推测结果改写文本以消除潜在泄漏线索,在多轮迭代中逐步提升匿名效果。然而,上述方法通常依赖大量算力对本地数据进行匿名化处理,难以部署在资源受限的用户终端设备上,严重影响了实际应用的可行性。此外,不同应用场景中用户的隐私信息类型可能存在差异,现有方法也难以有效利用分散在不同节点上的标注数据,以实现隐私识别任务的场景自适应与性能进一步提升。与上述方法不同,本文提出一种轻量化且具备场景自适应能力的隐私匿名化方法。通过知识蒸馏技术将大模型在隐私识别方面的能力有效迁移至轻量级本地模型中,并引入跨节点协同学习机制,实现对分散标注数据的高效利用。上述方法显著降低了计算和存储需求,使其更适合部署于资源受限的终端设备,同时根据不同场景下的隐私信息特点动态调整识别策略,在减少对集中标注数据依赖的同时,提升模型在多样应用环境中的泛化性与匿名化效果。

1.2　联邦学习

联邦学习核心思想是允许多个设备或数据持有方在不直接交换原始数据的前提下,协同训练全局模型。具体而言,每个客户端在本地执行模型更新,仅需将参数或梯度上传至中心服务器进行聚合,从而降低了隐私泄露风险。这一范式已在医疗诊断、金融风控及智能终端等多个隐私敏感场景中展现出应用潜

力。然而,现有的联邦学习方法在实际应用中仍存在多方面挑战,单纯的FedAvg[30]框架在实际应用中并不足以满足复杂场景需求,因此围绕性能优化与隐私安全保障,提出了大量改进方法。在性能方面,研究者主要关注如何提升联邦学习在异质数据环境下的收敛速度和稳定性。Google提出的FedAdam[31]旨在提升在异质数据环境下的训练效率和收敛性,通过在服务器端使用自适应优化器(如Adam)对客户端上传的模型更新进行加权平均,从而提高收敛速度和稳定性。Li等人[32]提出了FedProx算法,旨在解决联邦学习中系统异质性和统计异质性带来的挑战,通过在本地优化目标中引入一个近端项,使得算法在面对设备性能差异和数据分布不均时,能够更稳定地收敛。在隐私安全方面,差分隐私[33]成为联邦学习的主流保障手段。比如,McMahan等人[34]将用户级别的差分隐私机制引入联邦学习中,在经典的FedAvg[30]框架上,通过对每轮本地更新执行裁剪和高斯噪声添加,成功为LSTM语言模型提供(ϵ,δ)-DP保障,同时在数百万用户数据规模上仍收敛到与非私密版本近似的性能。Wei等人[35]提出NbAFL差分隐私方法,在客户端上传模型前加入高斯噪声以保证隐私,从理论上证明了该方法在不同噪声水平下可满足DP,并给出收敛性界限,揭示隐私与性能的权衡,同时指出增加客户端数量有助于收敛,并存在针对特定隐私水平的最优通信轮数与客户端规模。Girgis等人[36]则进一步提出CLDP-SGD算法,在联邦学习中结合本地差分隐私与通信效率,通过聚合客户端梯度并利用最优平均估计减少通信开销,同时借助客户端与数据子采样及Shuffled模型实现隐私放大,实现在凸损失下,CLDP-SGD能达到与集中式差分隐私相当的性能,而每轮仅需有限比特传输,兼顾隐私与效率。Mcmahan等人[14]在设备端联邦学习中通过引入基于在线优化的正则化更新机制与高效噪声聚合结构,有效平衡了差分隐私约束与模型可用性,在确保强隐私保护的同时,减少了噪声累积带来的性能退化,从而显著提升模型的训练稳定性与精度[37]。尽管现有方法在隐私保护与训练效率方面取得了一定进展,但仍存在关键局限。现有方法在模型训练过程中完全依赖于分散设备上的本地标注数据。当本地数据规模有限、标注样本稀缺或分布高度不均时,模型的收敛性与泛化能力将受到显著影响,难以在低资源场景中保持稳定性能。此外,当前方法在隐私保护与模型精度之间存在固有权衡,差分隐私噪声的引入虽能提升安全性,却常导致模型性能明显下降。虽然增加本地标注数据有助于缓解上述问题,但在隐私识别任务中,这一做法会进一步加剧潜在的隐私泄露风险,形成典型的"隐私困境"。为应对上述挑战,本文提出了一种新型隐私保护协同学习框架PrivateAI,在传统联邦学习范式的基础上实现了关键改进。不同于仅依赖分散标注数据的既有方法,PrivateAI首次将通用基础模型的隐私识别能力知识蒸馏与本地私有数据的协同学习结合,实现了隐私保护与模型性能的协同优化。PrivateAI首先利用基础模型推理行为中蕴含的通用隐私识别知识,经过安全提取生成轻量级蒸馏数据集。其次,PrivateAI构建了异构知识融合与对齐机制,有效整合基础模型知识与本地标注知识,实现多源知识的高效协同,从而显著提升模型在异构条件下的隐私识别能力。实验表明,PrivateAI能在差分隐私约束下最大化知识迁移效率,并在多个真实数据集上显著提升隐私保护成功率与识别精度。

2 方法

本节将介绍在生成式人工智能应用场景下针对隐私保护的PrivateAI方法。首先介绍本文所采用的差分隐私技术的定义,随后给出研究问题的形式化表述,最后系统介绍PrivateAI框架设计及其实现细节。

2.1 高斯差分隐私

局部差分隐私(Local Differential Privacy,LDP)方法[9]是分布式机器学习中广泛应用的隐私保护手段之一[38]。该方法通过对跨设备传递的数据添加随机扰动来保护其中的隐私信息,同时在一定程度上保持聚合数据的统计特性(例如均值)[36]。形式化地,若随机化算$\mathcal{R}(\cdot)$满足以下不等式,则其被认为符合(ϵ,δ)-LDP隐私保护要求:

$$\Pr\left[\mathcal{R}(X)\in\mathcal{X}\right]\leqslant e^{\epsilon}\cdot\Pr\left[\mathcal{R}(X')\in\mathcal{X}\right]+\delta \tag{1}$$

其中,X表示需要保护的隐私数据;$\mathcal{X}$表示输入数据的定义域;δ为松弛项;ϵ为隐私预算,用于控制隐私保护强度,ϵ值越小,隐私保护越严格。基于高斯噪声的随机化机制是实现LDP的常用技术,该机制$\mathcal{M}(\cdot)$实现(ϵ,δ)-LDP的条件可表述为

$$\mathcal{M}(X)=X+N(0,\sigma^2),\sigma^2=\frac{2\Delta^2\log\frac{1.25}{\delta}}{\epsilon^2},\quad \Delta^2=\max_{X,X'\in\mathcal{X}}\|X-X'\|^2 \tag{2}$$

其中,$N\left(0,\sigma^2\right)$表示均值为0、方差为σ^2的高斯分布;$\|\cdot\|^2$表示L2范数。在本文提出的PrivateAI框架中,采用该高斯机制对跨设备传输的中间表征或模型参数加以扰动,从而在协同训练过程中有效保护用户隐私,同时保持模型整体的学习效能。

此外,差分隐私提供攻击无关、可证明的统计隐私保证。一旦算法满足(ϵ,δ)-差分隐私,其输出分布在任意相邻数据集上的变化被严格上界,从理论上限

制了单个样本对模型行为的影响。基于这一性质,差分隐私对攻击者能力具有攻击无关性的防护效果,可统一覆盖多类隐私攻击,包括成员推断、属性推断以及模型反演等攻击范式。因此,相较于针对特定攻击设计的经验性防御机制,满足差分隐私的算法在更广泛的攻击场景下均具有可证明的安全保障。

2.2　问题定义

本节对本文研究的问题进行形式化定义。为了保护在线生成式人工智能应用中的用户隐私,需构建一个高精度的隐私识别模型 $\mathcal{J}(\cdot)$,部署于用户终端,在用户上传指令数据前识别潜在的隐私风险。令 Θ 表示模型 $\mathcal{J}$ 的参数。训练一个有效的隐私识别模型通常依赖以下两方知识来源。首先,是用于隐私识别任务的标注数据,该数据包含隐私信息与非隐私信息,以辅助模型区分两类内容。由于在实际应用中,标注任务数据分布于不同用户的设备上,本文假设部分用户在使用生成式人工智能应用时具有隐私保护需求,因而自愿参与协同学习训练隐私识别模型。令参与协同训练的用户集合为 $\mathcal{U}=\{u_i|i=1,2,\cdots,K\}$,其中 u_i 表示第 i 个用户,K 表示协作用户总数,用 $\mathcal{S}_u$ 表示用户 u 的本地标注数据集合。其次,另一重要的任务知识来源是基础大模型(如 DeepSeek),由于大模型编码了大量通用先验知识,在隐私识别任务上具备较强的零样本或少样本推理能力,能够进一步隐私识别模型的训练效果。在实际场景中,这些基础模型通常部署在云平台,仅能通过接口访问推理结果。本文假设在协同学习中可使用 H 个基础模型,第 i 个可用基础模型记为 $\mathcal{F}_i$,通过 $\mathcal{F}_i$ 给提供查询数据 x 可获得相应推理结果 $\mathcal{F}_i(x)$。此外,假设存在一个从公开渠道收集的无标注代理数据集 $\mathcal{P}=\{x_i|i=1,2,\cdots,E\}$,用于从基础模型中提取任务相关知识,其中 x_i 表示 $\mathcal{P}$ 中的第 i 个样本,E 表示无标注样本总量。本文还假设协同学习过程由一个中央服务器组织并协调多方参与。PrivateAI 框架的目标是在满足隐私约束的前提下,协同利用分散的标注数据与基础模型能力,构建隐私识别模型。

PrivateAI 采用迭代训练机制,协同利用分布式资源学习隐私识别模型,其整体流程如图 2 所示。在每一轮训练迭代中,PrivateAI 包含三个核心步骤:(1)标注数据知识提取;(2)基础模型知识提取;(3)层次化异质知识融合。具体而言,各用户设备首先从本地存储的标注数据集中提取与隐私识别相关的知识,并在施加适当隐私保护措施后,将知识表示上传至中央服务器。与此同时,服务器利用隐私无关的代理数据集,从多个基础模型中提取隐私识别任务相关的先验知识。随后,服务器对来自用户端和基础模型的异质知识进行对齐与融合,将整合后的知识注入隐私识别模型中。通过这一机制,分布式标注数据与云端基础模型能力得以在隐私保护的前提下被协同利用,有效提升模型性能。训练过程中,服务器将更新后的模型参数分发回各用户设备,迭代执行直至模型收敛。详细流程将在后文说明。

2.3　总体架构

在协同训练阶段完成后,经过充分优化得到的隐

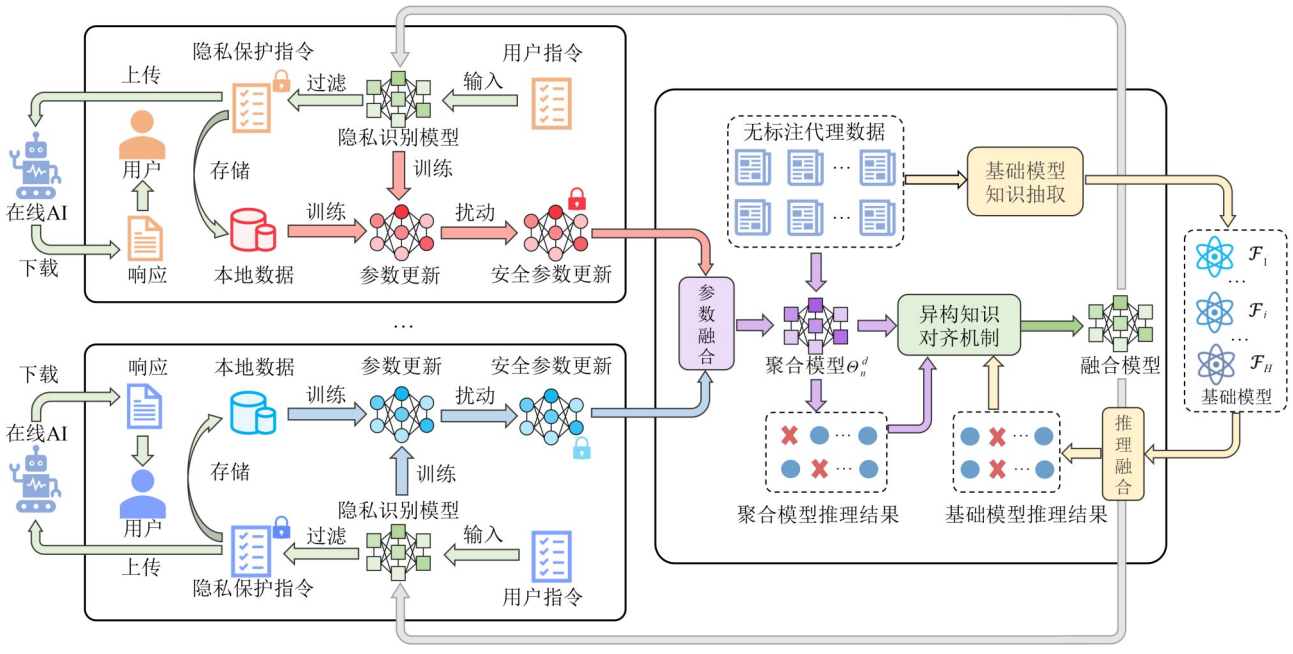


图 2　PrivateAI 方法框架

Figure 2　Framework of PrivateAI

私识别模型 $\mathcal{J}(\cdot;\ \Theta^*)$ 将被部署到各用户的设备中，以作为本地隐私保护框架的核心。具体而言，对于用户为调用生成式人工智能服务而输入的指令数据 T，该机制首先将其划分为若干细粒度指令单元，划分后的指令数据集合记为 $T=\{T_i|i=1,2,\cdots,G\}$，其中 T_i 表示第 i 个细粒度指令单元，G 表示划分单元的总数。随后，隐私识别模型对每个细粒度单元进行隐私信息的识别，并输出相应的隐私风险评估结果 $\{r_i|i=1,2,\cdots,G\}$，其中 $r_i=\mathcal{J}(T_i;\ \Theta^*)$ 表示对 T_i 的隐私风险评估。本地系统根据评估结果向用户提示可能存在的隐私风险，辅助其对原始指令进行修改或脱敏，最终生成一份具有隐私保障的指令数据版本，使用户能够安全地使用生成式人工智能服务，显著降低隐私泄露风险。

2.4 隐私保护协同学习

本节将介绍 PrivateAI 的隐私保护协同学习框架的具体细节。在第 n 轮训练迭代中，每个用户设备上的本地模型会首先与中央服务器维护的最新全局模型同步，用 Θ_n 表示第 n 轮初始化时的全局模型。第一阶段为标注数据知识提取。中央服务器从用户集合中随机采样一个子集 $\mathcal{U}_n$ 参与当前轮的协同学习。随后，每个被采样的用户 $u\in\mathcal{U}_n$ 在其本地数据上训练当前的隐私识别模型：

$$\begin{aligned}&\Theta_n^u=\Theta_n-\gamma\frac{\partial\mathcal{L}(\mathcal{S}_u;\ \Theta_n)}{\partial\Theta_n^u},\\&\mathcal{L}(\mathcal{S}_u;\ \Theta_n)=\frac{1}{|\mathcal{S}_u|}\sum_{(x,y)\in\mathcal{S}_u}l(y,\hat{y}),\\&\hat{y}=\mathcal{J}(x;\ \Theta_n)\end{aligned}\tag{3}$$

其中，Θ_n^u 表示局部优化后的参数；γ 为本地学习率；l 为损失函数。通过该过程，本地标注数据中与隐私识别相关的任务知识被提取到本地模型参数中。然而，由于知识载体（即本地模型 Θ_n^u）直接来源于用户本地数据，可能与用户隐私存在潜在关联，因此在与外部共享的过程中需要采取相应的隐私保护措施。本文采用基于高斯噪声扰动的局部差分隐私机制进行保护：

$$\begin{aligned}&\hat{\Theta}_n^u=h(\Theta_n^u;\ \theta)+n_n^u,\quad n_n^u\sim N(0,\sigma^2),\\&h(\Theta;\ \theta)=\frac{\theta}{\max(\theta,\|\Theta\|_2)}\cdot\Theta\end{aligned}\tag{4}$$

其中，$\hat{\Theta}_n^u$ 表示 Θ_n^u 加入隐私保护后的参数；$h(\Theta_n^u;\ \theta)$ 为归一化函数；θ 为参数的最大 L2 范数。本地设备 u 将隐私保护后的知识载体 $\hat{\Theta}_n^u$ 发送至中央服务器用于协同学习。需要注意的是，当 $\sigma^2=8\theta^2\log\frac{1.25}{\delta}/\epsilon^2$ 时，隐私信息可在 (ϵ,δ)-LDP 的约束下得到有效保护。

第二阶段为基础模型知识提取。强大的基础模型通常在模型平台上（如 OpenAI）部署，其参数往往无法直接获取。因此，PrivateAI 的核心思路是通过基础模型在代理数据上的推理结果来提取相应的隐私识别知识。具体而言，对于给定的第 k 个基础模型 $\mathcal{F}_k$，本文首先设计了一套指令策略 $p_k(x)$ 提示模型 $\mathcal{F}_k$ 识别数据 x 中的隐私敏感内容。例如，在 ChatGPT 模型中，该指令策略可通过将预设提示与目标数据 x 拼接的方式实现。随后，对于代理数据集 $\mathcal{P}$ 中的每个数据 x_i，服务器将其对应的指令化输入 $p_k(x_i)$ 发送至基础模型 $\mathcal{F}_k$，获得其隐私识别推理结果 f_j^k。这样即可得到一组由基础模型 $\{\mathcal{F}_k|k=1,2,\cdots,H\}$，基于不同代理数据生成的隐私识别推理结果 $F_k=\{f_j^k|j=1,2,\cdots,E\}$，这些结果可作为编码了不同基础模型隐私识别能力的知识载体。

第三阶段为层次化异质知识融合。在此阶段，服务器需将分将从分布式数据与基础模型提取的隐私识别知识融合至全局模型 Θ_n，以完成本轮协同学习。然而，不同来源（数据与模型）的知识被编码在不同的知识载体中（模型参数与模型推理结果），导致相应的知识难以直接融合。为此，本文提出了一种层次化的异质知识融合方法，首先融合来自不同来源的同类型载体中的知识，并进一步整合编码在不同载体的异质知识。具体来说，PrivateAI 首先对于以本地模型参数作为载体所编码的分布式标注数据中的知识，通过基于自适应动量[39]的模型平均来实现融合：

$$\begin{aligned}&\Theta_n^d=\Theta_n+\eta\frac{G_n}{\sqrt{\kappa_n+\tau_0}},\\&G_n=\tau_1G_{n-1}+(1-\tau_1)\nabla_n,\\&\kappa_n=\tau_2\kappa_{n-1}+(1-\tau_2)\nabla_n^2,\\&\nabla_n=\frac{1}{|\mathcal{U}_n|}\sum_{u\in\mathcal{U}_n}\hat{\Theta}_n^u-\Theta_n\end{aligned}\tag{5}$$

其中，Θ_n^d 表示融合后的分布式数据知识载体；η 为更新参数；G_n 与 κ_n 分别为一阶与二阶动量；∇_n 为局部更新的平均值；τ_0、τ_1、τ_2 为超参数。其次，对于以模型推理结果作为载体所编码的隐私识别知识，PrivateAI 采用加权平均知识融合，以降低单一基础模型的有偏推理结果对模型更新的影响：

$$\bar{F}=\{\bar{f}_j|j=1,2,\cdots,E\},\quad\bar{f}_j=\frac{1}{H}\sum_{k=1}^{H}f_j^k\tag{6}$$

其中，$\bar{f}_j$ 表示在代理数据 $x_i\in\mathcal{P}$ 上融合后的知识。

进一步地，Private AI 需要融合两类异质知识载体 Θ_n^d 与 $\bar{F}$ 中所表征的任务知识，并将其更新到模型参数中。由于模型参数与模型推理数据的知识表达形式具有显著差异，难以直接融合。为此本文提出了异构知识融合机制以解决上述挑战。首先，PrivateAI 将基于模型参数的任务知识转化为基于推理结果的任务知识。具体来说，PrivateAI 方法利用基于参数 Θ_n^d

的隐私识别模型 $\mathcal{J}(x;\Theta_n^d)$ 对代理数据 $\mathcal{P}$ 进行预测，得到模型推理结果 $\hat{F}=\{\hat{f}_j|j=1,2,\cdots,E\}$，其中 $\hat{f}_j$ 表示对代理数据集 $\mathcal{P}$ 中第 j 条数据的推理结果。其次，PrivateAI 进一步在模型推理表征空间，对来自分散标注数据以及基础模型的任务知识进行融合。通常情况下，来自不同来源的任务知识可能存在冲突，因此需要在融合前进行对齐从而提升学习性能。为此，PrivateAI 平均了两种来源的推理结果，以对齐来自分布式数据和基础模型的知识：$f_j=\frac{1}{2}\bar{f}_j+\frac{1}{2}\hat{f}_j$。此外，为了解决严重的知识冲突问题，PrivateAI 在后续学习中忽略了显著不一致的推理结果：$\mathcal{P}_n=\left\{(x_j,f_j)|x_j\in\mathcal{P},f_j>\mu,j=1,2,\cdots,E\right\}$，以避免对隐私识别模型的训练带来干扰，其中 μ 表示过滤不一致知识的阈值。通过这种方式，$\mathcal{P}_n$ 可对齐并表示来自分布式数据和基础模型的隐私识别任务知识。最后，PrivateAI 在 $\mathcal{P}_n$ 上训练模型 Θ_n^d，即可将对齐后的知识注入全局模型，并得到更新后的全局模型 Θ_{n+1}。PrivateAI 而后将更新后的全局模型 Θ_{n+1} 分发至各用户设备，用于下一轮训练迭代，直至模型收敛。PrivateAI 的详细流程见算法 1。此外，为进一步加速模型收敛效率，PrivateAI 在实际执行时可以以 I_{kd} 为间隔进行知识蒸馏，I_{kd} 取值范围为 $[1,\infty)$，其中 $I_{kd}=1$ 表示每轮模型聚合均进行蒸馏数据训练，$I_{kd}=10$ 则表示每 10 轮聚合仅执行 1 次蒸馏训练，而 $I_{kd}=\infty$ 则表示模型不执行蒸馏训练。通过上述方式，我们在大幅降低模型额外计算开销的同时，有望保持训练精度。

2.5　差分隐私安全分析

在 PrivateAI 方法中，各节点之间仅交换模型参数而非原始数据。因此，只需保证参数交换过程满足 (ϵ,δ)-差分隐私，即可依据差分隐私的后处理不变性，确保对原始训练数据同样满足 (ϵ,δ)-差分隐私。为此，在参数交换之前，我们首先将模型参数的 l_2-范数裁剪至不超过阈值 θ（见式(4)），随后对裁剪后的参数添加高斯噪声。由于任意两个相邻数据集对应的参数均被裁剪至半径为 θ 的球内，其 l_2-敏感度可被上界为 $\Delta_2=\max\|\Theta-\Theta'\|_2\leqslant 2\theta$。因此，根据高斯机制的隐私保证，当所添加噪声的方差满足 $\sigma^2\geqslant\frac{8\theta^2\ln\left(\frac{1.25}{\delta}\right)}{\epsilon^2}$，参数交换过程即可满足 (ϵ,δ)-差分隐私。进一步地，由于模型参数是原始训练数据的确定性或随机后处理结果，PrivateAI 方法整体上实现了对原始训练数据的 (ϵ,δ)-差分隐私保护。此外，尽管大模型训练过程中通常涉及大规模、高频词的参数交换，可能在理论上引入隐私预算累积的风险，但已有研究表明[30]，可通过匿名通信机制对不同轮次的参数更新进行去关联，单轮参数交换所满足的 (ϵ,δ)-差分隐私保证可以自然推广至整个训练过程。

算法 1　PrivateAI 工作流

输入：用于协同学习的本地设备及其私有数据集 $\{(u,S_u)|u=1,2,\cdots,K\}$，用于协同学习的基础模型 $\{\mathcal{F}_j|j=1,2,\cdots,H\}$，用于协同学习的隐私无关代理数据集 $\mathcal{P}$，PrivateAI 使用的超参数 $\gamma,\theta,\delta,\epsilon,\eta,\mu$ 等，初始的隐私识别模型 $\mathcal{J}(\cdot;\Theta)$

输出：收敛后的隐私识别模型

WHILE Θ_n 未收敛时

//伪代码注释：第 n 轮的标注数据知识抽取

[S]: 采样部分本地设备用于进行模型训练，得到集合 $\mathcal{U}_n$

FOR $u\in\mathcal{U}_n$

//伪代码注释：执行针对来自设备 u 的标注数据的知识抽取

[u]: 基于本地标注数据 S_u 训练当前的模型参数 Θ_n，得到更新的本地模型 Θ_n^u

[u]: 上传更新的本地模型 Θ_n^u 至中央服务器

END FOR

FOR $k=1,2,\cdots,H$

//伪代码注释：第 n 轮的基础模型知识抽取

[S]: 基于代理数据 $\mathcal{P}$，生成面向模型 $\mathcal{F}_k$ 的查询指令 $\mathcal{P}_k$

[S]: 使用指令化的代理数据 $\mathcal{P}_k$ 查询基础模型 $\mathcal{F}_k$

[S]: 得到推理结果 F_k

END FOR

//伪代码注释：第 n 轮异构知识融合

[S]: 融合来自各设备的标注数据知识 $\{\Theta_n^u|u\in\mathcal{U}_n\}$，聚合本地模型，得到统一模型 Θ_n^d

[S]: 融合来自不同基础模型的知识 $\{F_k|k=1,2,\cdots,H\}$，聚合推理结果，得到统一推理结果 $\bar{F}$

[S]: 将表征标注数据编码知识的载体从模型参数 Θ_n^d 转换为推理结果 $\hat{F}$

[S]: 对齐来自标注数据知识和基础模型的任务知识，得到 F

[S]: 过滤不同来源的冲突知识，仅保留对齐的知识数据，得到 $\mathcal{P}_n$

3　结果

3.1　实验设置

本研究基于两个真实场景的数据集开展了实证分析，以评估不同隐私保护机制的有效性。第一个数据集为 EMR-Smoke[40]，基于电子病历中关于患者吸烟状态的记录构建。该数据集中可能包含多种受保护健康信息，例如患者 ID、姓名和电子邮件地址[41]等。在线人工智能应用的隐私保护场景中，终端用户可能会将此类病历上传至在线平台以寻求专业医疗服务，因此其中的隐私信息必须得到严格保护。第二个数据集为 EMR-Cardiac[42]，由心血管疾病相关的电子病历构成。与前一个数据集类似，EMR-Cardiac 中的病历同样包含需要采取隐私保护措施的敏感信息。

为模拟实际应用中训练数据分布于多个设备的情况，将每个数据集的训练数据划分为多个簇。同时，真实环境下本地数据往往具有非独立同分布（non-IID）的特性，这是分布式学习中的普遍挑战。为此，在划分时引入了非 IID 采样策略，并通过参数 α 控制数据分布的不平衡程度，越小的 α 值意味着设备间数据差异越明显。此外，PrivateAI 在模型训练中需要借助基础模型提供协同知识支持。本实验使用 DeepSeek-V3 与 GPT-4o 为隐私保护任务提供知识。同时，PrivateAI 还需依赖无标签代理数据来提取基础模型知识。为此，本文从医学文本数据集 MT 中收集医学文本，构建了 PrivateAI 的代理数据集。隐私保护的成功性以模型识别隐私信息的准确程度为衡量标准。当隐私识别模型能够以绝对准确率检测出隐私信息时，隐私威胁即可消除。因此，本研究采用隐私保护成功率（Protection Successful Rate，PSR）作为核心评估指标，其计算基于模型对隐私信息的识别准确率。

PrivateAI 的超参数设置总结如下。在每轮迭代中，随机抽取 80% 的可用设备参与协同学习。在本地知识提取过程中，每个被选中的设备利用其本地数据对局部模型进行训练，优化器采用 Adam[39]。Transformer 模型的本地学习率 γ 设置为 10^{-4}，BERT 与 RoBERTa 模型则设为 10^{-5}。在知识交换过程中，通信模型 θ 的最大 L2 范数设置为 1.0，隐私松弛系数 δ 设置为 $\frac{1}{|s|}$，其中 $\mathcal{S}=U_{u\in\mathcal{U}}\mathcal{S}_u$ 表示所有参与设备上的训练数据集合。隐私预算设定为 ϵ，其中 $\epsilon\in\{0.01,0.1,1,10^{0.5},10^{1},10^{1.5},10^{2.0},10^{2.5},10^{3.0}\}$，以评估不同隐私保护强度下的方法性能。在聚合局部模型所携带的任务知识时，知识更新率 η 在 Transformer 上设置为 10^{-4}，其他模型上为 10^{-5}。其余聚合超参数采用常见文献[39]中的设定，即 τ_0、τ_1、τ_2 分别设为 10^{-8}、0.9、0.999。在聚合不同载体所承载的异质知识时，知识对齐阈值 μ 设为 0.9。训练全局模型时，同样采用 Adam 优化器，其中 Transformer 的学习率设置为 10^{-4}，BERT 与 RoBERTa 则为 10^{-5}。

本文对 PrivateAI 与三种具有代表性的隐私保护协同学习方法进行对比分析，以验证其在学习隐私识别模型中的优越性。对比方法包括：

（1）FedAvg[30]。在分散数据上训练隐私识别模型，通过对来自不同用户设备的本地优化模型参数进行平均，不断迭代更新共享模型。

（2）FedAdam[31]。一种基于自适应矩估计的分布式学习框架，用于训练共享模型。

（3）FAFL[43]。一种基于自适应矩阵优化的分布式学习框架，用于训练共享模型。

（4）RobustDPF[13]。将 Rényi 差分隐私机制引入联邦学习框架中，通过在本地节点上传模型参数前注入噪声来实现扰动，从而有效保护训练过程中本地数据的隐私安全。

（5）DP-FTR[14]。通过引入正则化优化与树形聚合机制，在保证强差分隐私的同时有效抑制噪声累积，提升联邦学习模型的训练稳定性与精度。

PrivateAI 与上述基线方法均需要在用户设备之间共享学习变量以实现协同学习。然而，现有研究指出，这类变量（如模型参数）的通信过程可能带来潜在隐私风险[44]。为保障协同学习过程中的隐私安全，在上述方法中引入本地高斯差分隐私技术来保护通信数据。隐私保护水平由隐私预算参数 $\epsilon(\epsilon>0)$ 控制，ϵ 值越小，隐私保护强度越高。在实际应用中，用户设备上的计算资源通常有限，难以训练和部署诸如大型语言模型等计算开销巨大的模型。因此，本文选择了三种参数规模较小或适中的基础机器学习模型作为隐私识别模型，包括：

（1）Transformer[45]。一个参数随机初始化的单层 Transformer 网络。

（2）BERT[21]。一个包含 12 层 Transformer 网络的预训练模型，总参数量约 1.1 亿。

（3）RoBERTa[46]。一个基于预训练的 24 层 Transformer 网络模型，参数量约 3.55 亿。

3.2 算法性能评估

本研究基于两个真实场景的数据集开展了实证分析，以评估不同隐私保护机制的有效性。第一个数据集为 EMR-Smoke[40]，基于电子病历中关于患者吸烟状态的记录构建。该数据集中可能包含多种受保护健康信息，例如患者 ID、姓名和电子邮件地址[41]等。本节在不同的隐私保护约束下，对多种协同学习方法训练的隐私识别模型进行性能评估。每组实验均重复 15 次，在图 3 中呈现平均结果。首先，在严格的差分隐私约束下，基线协同学习方法的表现显著下降。例如，在 EMR-Cardiac 数据集上，以 $\epsilon=1$ 的隐私级别训练 Transformer 模型时，FAFL 的隐私保护成功率仅约 40.5%。这一现象源于差分隐私机制需向通信的本地模型中注入噪声以保证隐私，而隐私约束越严格，扰动越强。由于基线方法完全依赖分散数据所承载的知识来训练模型，其性能极易受到本地模型退化的影响。第二，部分基准方法（Robust-DPFL、DP-FTRL）通过引入更高效的差分隐私机制显著提升了模型的学习性能。例如，当差分隐私预算参数 $\epsilon=10$ 时（EMR-Smoke 数据集），采用基准算法 FedAdam 训练基于 RoBERTa 的隐私识别模型仅能达到 40.5% 的准确率，而使用 DP-FTRL 方法可将准确率提升至

56.5%,验证了 DP-FTRL 方法在差分隐私效率方面的优势。相比之下,本文提出的 PrivateAI 方法在相同条件下实现了 87.3% 的训练准确率,仍然显著超越了现有 SOTA 方法。该性能提升主要得益于 PrivateAI 不仅能够利用分散数据中包含的知识,还能将基础模型中的先验知识注入隐私识别模型。通过融合来自多源的异质知识,PrivateAI 对数据知识退化表现出更强的鲁棒性,从而保障了协同学习的有效性。尤其值得注意的是,在所有隐私安全级别下,PrivateAI 始终优于基线方法。例如,在 EMR-Cardiac 数据集上,以 $\epsilon=1$ 训练 RoBERTa 时,PrivateAI 的 PSR 达到 86.5%,相较于最优基线方法提升了 47.0 个百分点。这一结果表明,来自基础模型的知识与分散数据的知识具有互补性,而二者的融合能够带来最优性能。上述结果充分验证了 PrivateAI 在兼顾隐私保护与模型有效性方面的显著优势。此外,PrivateAI 也存在在一定隐私预算($\epsilon\in[1,1\,000]$)范围内性能变化较小的现象,即当隐私预算从 1 000 逐步降低至 1 时,且在利用 PrivateAI 对预训练语言模型(如 BERT、RoBERTa)进行构建隐私识别模型时,模型性能整体保持相对稳定。上述现象主要源于以下两个方面的原因:

(1)本文引入了基于大模型的知识蒸馏机制。在隐私预算较大时,对应的差分隐私噪声扰动相对较弱,蒸馏过程能够有效吸收并平滑由隐私噪声引入的随机性,从而显著缓解噪声对模型优化过程的干扰,使模型性能在较宽的隐私预算范围内保持稳定。

(2)预训练语言模型本身已在大规模语料上学习到丰富的先验知识,使其在微调阶段对差分隐私噪声具有更强的鲁棒性,从而进一步减弱了适中的噪声扰动对最终性能的影响。

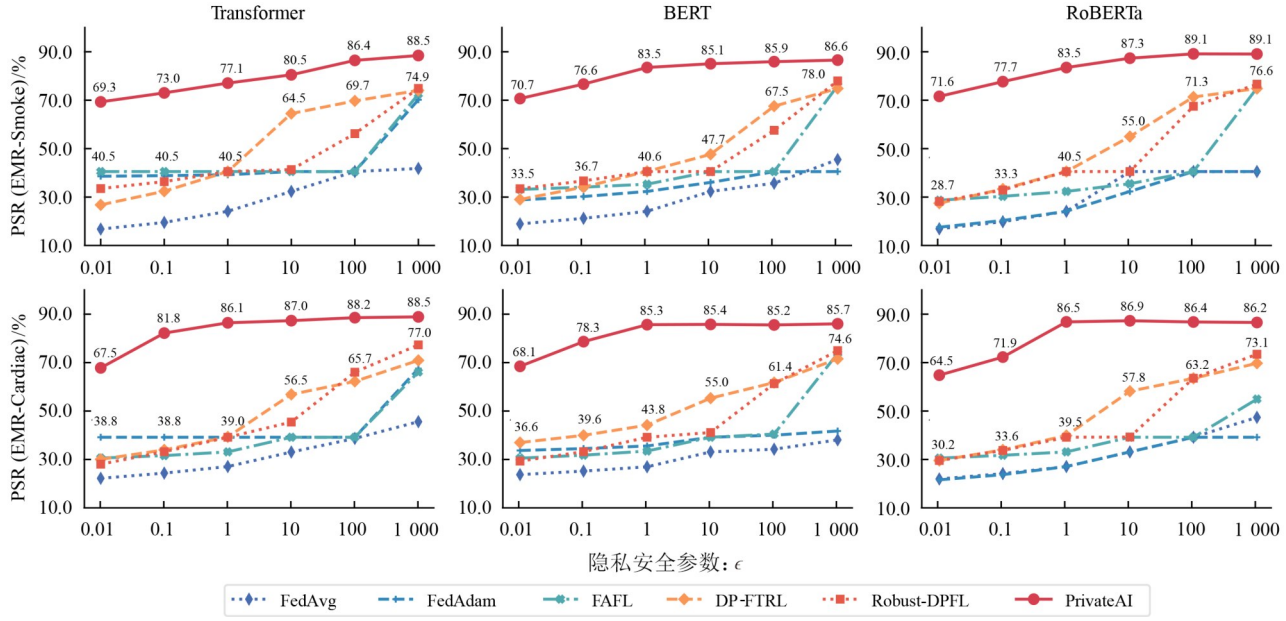


图 3　不同方法在学习隐私识别模型时的性能对比

Figure 3　Performance comparison of different methods in learning privacy recognition models

3.3　算法性能评估

为了更加直观地反映算法的隐私安全,本次修订选取了当前最具代表性的面向人工智能模型的隐私攻击方式,即成员推理攻击,对算法的隐私风险进行量化评估。该攻击建立在较强的攻击者先验假设之上,即假设攻击者能够获取一批可能包含训练数据的候选样本集合,但并不明确其中哪些样本实际参与了模型训练。其核心目标在于判定给定数据样本是否被用于模型训练,是衡量模型隐私泄露风险的标准评估手段之一。根据审稿专家的建议,我们在修订稿中进一步补充了基于成员推理攻击[47]的隐私安全性分析与实验验证。实验结果表明(见图 4),当 $\epsilon\leqslant 10$ 时,成员推理准确率接近随机水平,训练数据的隐私性得到了非常有效的保护;当 $\epsilon=1\,000$ 时,成员推理成功率仅为 0.6 左右,表明存在一定的隐私泄漏风险,但相较于无隐私保护情形已显著降低。上述实验结果,证明了在适当的隐私预算下($\epsilon\leqslant 10$),我们的 PrivateAI 方法对训练数据隐私安全的保障。

3.4　数据异质性的影响

本节对不同方法在数据异质性条件下的鲁棒性进行分析。本节首先评估了 PrivateAI 与代表性基线方法 FedAdam 在任务数据异质性场景下的性能。这里的任务数据异质性指设备间本地数据分布差异,具

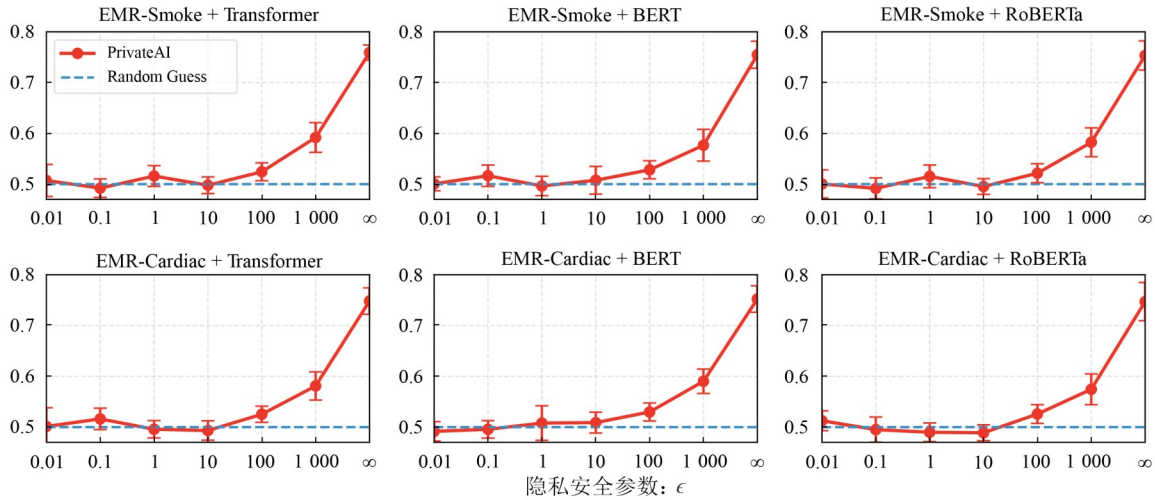


图4 基于成员推理攻击的训练数据隐私风险评估结果

Figure 4 Results of privacy risk evaluation on training data based on membership inference attacks

体划分为三种情况：极端异质性（$\alpha=0.1$）、强异质性（$\alpha=1.0$）和弱异质性（$\alpha=10.0$）。此外，PrivateAI在训练中除依赖带标注的任务数据外，还需利用代理数据来提取基础模型的知识。带标注任务数据与代理数据之间的差异（以下简称“迁移数据异质性”）同样会影响PrivateAI的性能。因此，本节进一步在不同的迁移数据异质性水平β下对PrivateAI进行了评估，包括0.1、1.0和10。实验结果如图5所示。

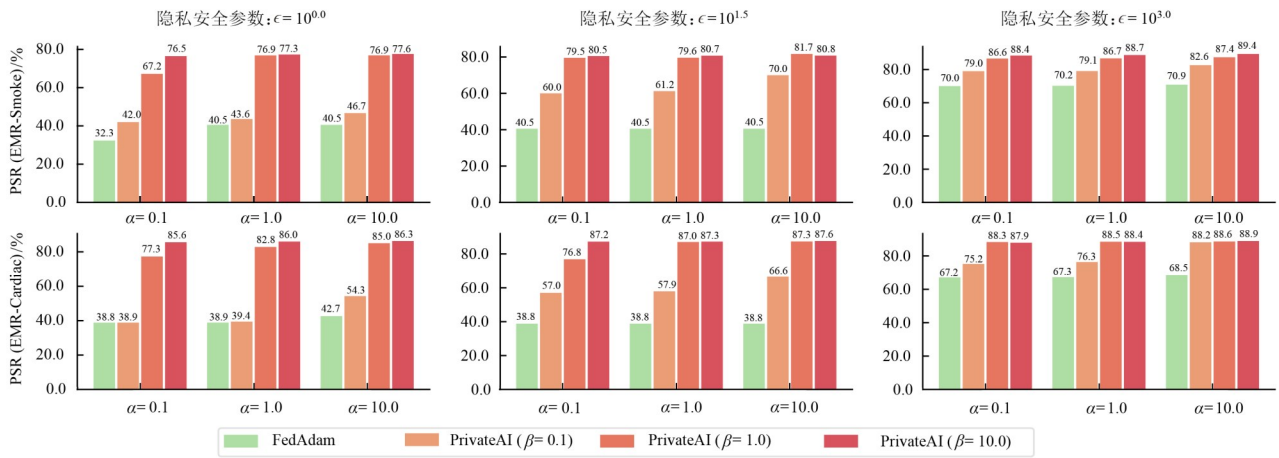


图5 在不同数据异质性条件下PrivateAI与FedAdam的有效性比较

Figure 5 Comparison of the effectiveness of PrivateAI and FedAdam under different data heterogeneity conditions

首先，在中等迁移数据异质性下，PrivateAI对任务数据异质性表现出较强鲁棒性。例如，在隐私安全水平$\epsilon=1$、弱迁移数据异质性（$\beta=10.0$）条件下，在EMR-Cardiac数据集上的实验表明，PrivateAI的PSR在任务数据异质性α从0.1~10.0的范围内保持稳定。这表明在该条件下，代理数据提取的基础模型知识能够与分布式数据知识形成互补。通过对多源知识进行对齐，PrivateAI有效优化隐私识别模型的训练过程。其次，当分布式数据能够提供充分任务知识时，PrivateAI对迁移数据异质性同样展现出较良好的适应性。例如，在EMR-Cardiac数据集上，当隐私预算$\epsilon=10^3$且任务数据异质性$\alpha=10.0$时，PrivateAI在弱迁移数据异质性（$\beta=10.0$）下的性能相比于极端迁移异质性（$\beta=0.1$）仅下降了0.7个百分点。这种鲁棒性归因于其能够从分布式数据中准确提取任务知识，从而有效抵消由迁移异质性引入的噪声。这些结果表明，PrivateAI对任务数据与代理数据之间的异质性具有鲁棒性，进一步验证了其在真实应用场景中的可行性。最后，PrivateAI在不同实验设置下均显著优于基线方法。尤其是在高度挑战性的场景下，如严格的隐私预算（$\epsilon=1$）和极端数据异质性（$\alpha=0.1$，$\beta=0.1$），PrivateAI在EMR-Smoke数据集上的PSR相较最优基

线仍提升9.7个百分点。这表明即便在真实场景中面对苛刻隐私约束与严重数据异质性,PrivateAI依然能够保持显著的性能优势。

3.5 消融实验

本节进行消融实验评估PrivateAI所引入的知识融合机制的有效性。本节基于Transformer模型(如图6所示),对PrivateAI与其消融版本进行比较。(1)w/o model knowledge:仅依赖自训练策略[48]在代理数据上训练共享全局模型,不引入任何来自基础模型的知识。(2)w/o knowledge alignment:直接利用基础模型的预测结果在代理数据上训练共享全局模型,但不与分布式任务数据的知识进行对齐。结果表明,首先PrivateAI始终优于w/o model knowledge,说明基础模型在为隐私识别任务提供丰富先验知识方面的关键作用,而仅依赖代理数据不足以支撑任务知识的获取。其次,尽管w/o knowledge alignment尝试将基础模型的任务知识融入隐私识别模型中,但其仅在w/o model knowledge的基础上带来了微弱的性能提升。结果表明,融合任务数据与基础模型之间的异构知识具有一定挑战。若缺乏有效的知识对齐机制,w/o knowledge alignment难以显著提升模型性能。第三,PrivateAI大幅提升了w/o knowledge alignment的效果,并显著优于其两个消融版本。例如,在EMR-Smoke数据集上、隐私预算设为$\epsilon=10^{1.5}$的情况下,PrivateAI在利用异构任务知识的效率上相较于消融版本最高提升47.3个百分点。这一优势得益于分布式数据与基础模型知识的有效对齐,并在训练过程中充分发挥二者的互补性,从而增强隐私识别模型的学习效能。这些结果凸显了PrivateAI在利用分散客户端的异构资源以构建更加高效的隐私识别模型方面的显著潜力。

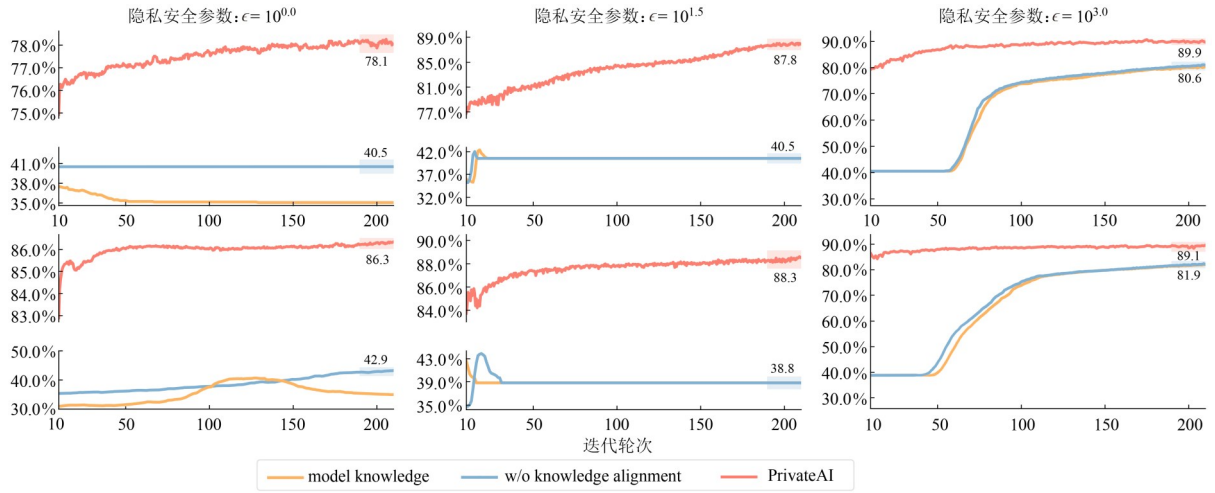


图6　PrivateAI的消融实验

Figure 6　Ablation study of PrivateAI

3.6 代理数据集规模分析

本节进一步探究代理数据规模对PrivateAI的影响,并在不同规模的任务数据条件下展开评估,同时以FedAdam作为对照方法,如图7所示。实验结果表明,随着代理数据量的增加,PrivateAI的性能快速提升,并在达到最优水平后趋于稳定。这是因为在代理数据不足时,基础模型任务知识无法充分迁移;而当代理数据过多时,基础模型容量成为瓶颈,从而限制了隐私识别模型的进一步优化。这一发现表明,无需收集海量代理数据来增强PrivateAI的性能,适量代理数据(如5 000条)足以使PrivateAI显著超越基线方法。其次,当标注任务数据稀缺时,代理数据带来的增益尤为突出。例如,在仅有$L=100$条标注任务数据,在EMR-Cardiac数据集上以隐私预算$\epsilon=1$进行实验时,PrivateAI的性能随代理数据增加从67.4%提升至86.0%。额外代理数据有效促进了基础模型知识的迁移,从而缓解了标注数据不足带来的性能下降。综上所述,PrivateAI能较好地适应真实场景中的数据约束,在仅依赖适量代理数据补充的情况下,即使标注任务数据有限,仍可保持优越性能。

3.7 计算效率对比

本节对多种方法在设备端与中心服务器端的开销进行了评估,从而消除在计算效率方面的潜在顾虑。在用户设备端,设模型参数数量为d,本地样本数为S。由于PrivateAI与基线方法均仅在本地设备上独立模型训练,故其计算复杂度相同,可表示为$\mathcal{O}(Sd^3)$,其中$\mathcal{O}(d^3)$表示在单个数据样本上训练模型的计算复杂度。在服务器端,基线方法仅执行模型聚合操作,计算复杂度为$\mathcal{O}(Kd)$。相比之下,PrivateAI除聚合本地模型外,还需整合基础模型的推理结果并

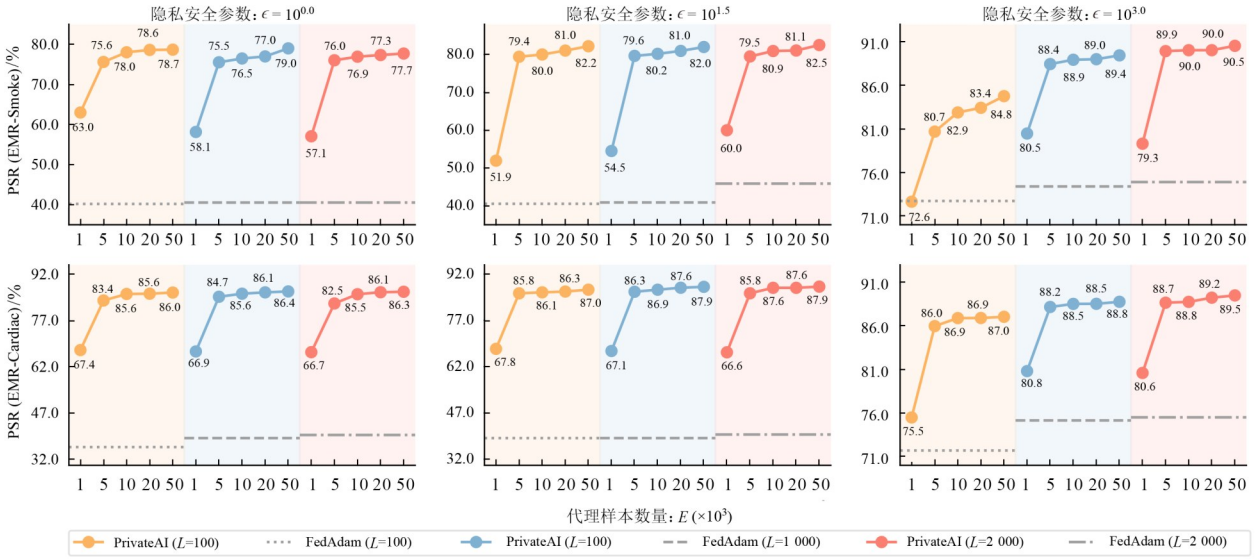


图7 代理数据规模对PrivateAI的影响

Figure 7 Impact of proxy data size on PrivateAI

在代理数据上进一步训练全局模型。其额外计算复杂度包括 $\mathcal{O}(HP)$ 与 $\mathcal{O}(Pd^3)$，其中 P 表示代理数据规模。由此，PrivateAI在服务器端的总体计算复杂度可表示为 $\mathcal{O}(Kd+PH+Pd^3)$。由于该复杂度的主要部分由 $\mathcal{O}(Pd^3)$ 主导，因此整体计算复杂度可近似为 $\mathcal{O}(Pd^3)$。需要指出的是，这部分额外开销主要来自全局训练。在实际应用中，隐私识别模型多为小型或中等规模，使得训练开销在现有服务器硬件条件下完全可接受。综上，PrivateAI不会增加本地设备的计算负担，其在服务器端的额外开销也处于合理可控范围。不同方法的计算复杂度对比如表1所示。此外，在实际的高频聚合场景中，本文提出的PrivateAI方法可以通过引入“低频蒸馏”机制，仅在部分模型聚合轮次中执行基于蒸馏数据的额外训练，从而在充分发挥大模型知识对分布式训练模型增强作用的同时，有效避免整体训练开销的显著增加。低频蒸馏的执行间隔为 I_{kd}，其取值范围为 $[1,\infty)$。我们进一步在实验中给出了不同PrivateAI的额外计算开销与算法收敛曲线（图8）。实验结果表明，适当增大基于基准模型知识的知识蒸馏间隔（如增加至每40轮对齐1次），可以在大幅度降低额外训练代价的同时改善模型的训练效果，最终，PrivateAI方法即便仅在少量采样聚合轮次中进行蒸馏训练，也能在极小的额外计算成本下显著提升模型的训练效果与收敛稳定性。

3.8 知识对齐超参数影响分析

本文提出的PrivateAI方法通过在代理数据集上对基础大模型中蕴含的通用隐私识别知识进行蒸馏，构建轻量化的蒸馏知识数据集，以增强分布式隐私训练的效果。由于本地标注数据与基础模型蒸馏所得

表1 不同方法计算复杂度在设备端与服务器端的对比

Table 1 Comparison of computation complexities on devices and servers

方法	设备端	服务器端
FedAvg	$\mathcal{O}(Sd^3)$	$\mathcal{O}(Kd)$
FedAdam	$\mathcal{O}(Sd^3)$	$\mathcal{O}(Kd)$
FAFL	$\mathcal{O}(Sd^3)$	$\mathcal{O}(Kd)$
PrivateAI	$\mathcal{O}(Sd^3)$	$\mathcal{O}(Kd+PH+Pd^3)$

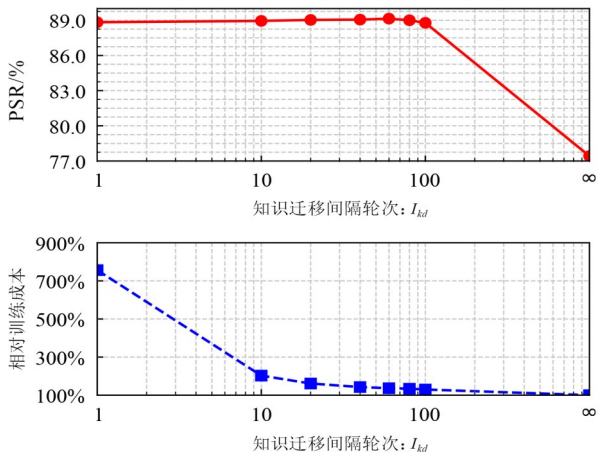


图8 PrivateAI方法对超参数 I_{kd} 的灵敏性分析

Figure 8 Sensitivity Analysis of the hyperparameter I_{kd}

隐私识别知识在表征形式与分布空间上存在显著差异（分别编码于本地模型参数与蒸馏数据集中），PrivateAI引入了一种基于代理数据集的表征对齐与冲突过滤机制，并通过超参数 μ 控制冲突知识的过滤阈值。该超参数 μ 的取值介于 $[0,1]$ 之间，其中 μ 值越大，表示仅当基于本地标注数据的预测结果 f 和基于大模型知识的预测结果 f 高度一致时，蒸馏样本才会

被保留。为分析该超参数对算法分析的影响,我们继续采用了2个数据集及2种典型的基准模型(Transformer、BERT)进行实验分析,图9给出了PrivateAI方法在不同超参数μ下的性能。结果表明,μ的取值在不同数据集与模型架构下具有较好的稳定性与普适性,当超参数μ取值介于0.8~0.9之间时,PrivateAI算法普遍取得了最优性能。在本文中,μ设置为0.9,用以过滤掉冲突较为严重的知识样本,从而避免错误知识干扰模型训练,同时保留具有一定差异的样本,以提升模型的收敛效率与泛化能力。

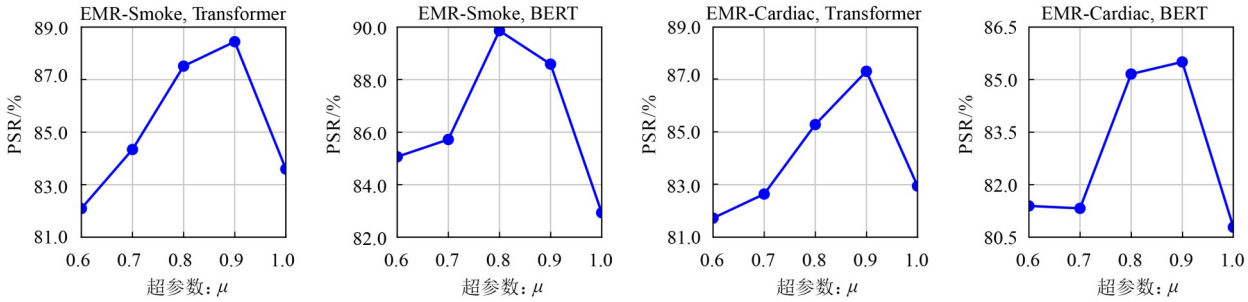


图9　PrivateAI方法对超参数μ的灵敏性分析

Figure 9　Sensitivity analysis of the hyperparameters μ

3.9　本地设备数目的影响

本节进一步评估了本地设备数量对PrivateAI性能的影响,并在两种不同条件下进行了实验。首先,在每个本地设备所持有的训练样本数量固定(记为S)的条件下,分别就$S=10$、20两种情况评估设备数量K的影响(如图10上部分)。实验结果表明,随着设备数量的增加,PrivateAI的性能逐步提升。这一现象可归因于更多设备提供了更大规模的训练数据,从而提升了模型学习效果,也再次印证了设备间协同对于构建高效智能系统的重要性。其次,在总可用训练样本在总数量L固定时,分别就$L=1\,000$、2 000的两种情况分析设备数量K的影响(如图10下部分)。结果显示,当K较小时,模型性能随设备数量增加而提升,这是由于更多设备能够更好地抵消LDP噪声在聚合阶段的影响。当K进一步增强时,若总样本量充足($L=2\,000$),PrivateAI的性能仍然能够持续提升;但在总样本量有限($L=1\,000$)时,性能则会出现小幅下降,这是由于单设备可用样本量过少导致的训练不充分。综上所述,设备数量的变化仅对PrivateAI的性能产生轻微波动,该稳定性得益于基础模型先验知识的引入,其在任务数据知识退化的条件下显著增强了模型训练的鲁棒性。

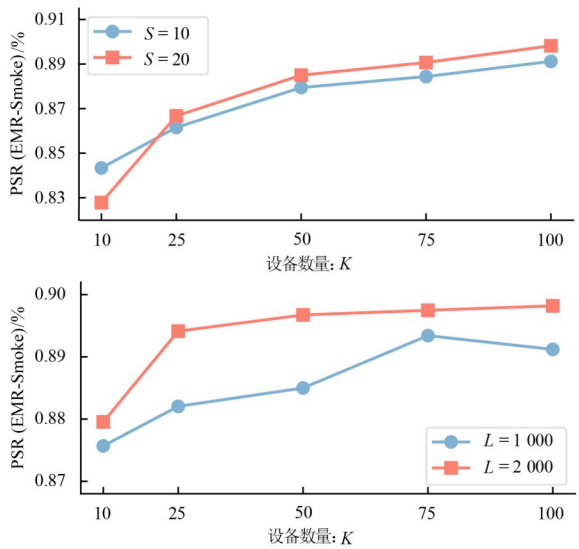


图10　在固定本地、全局样本规模下设备数量对PrivateAI的影响

Figure 10　Impact of the device sizes on PrivateAI under a given local sample size or global sample size

3.10　代理数据集分布影响的初步分析

PrivateAI的核心机制在于从基础大模型的推理行为中提炼隐私识别知识,并基于代理数据集P进行蒸馏,从而实现知识的轻量化迁移与分布式隐私模型的增强。代理数据集在其中起到“知识载体”的作用,其分布特征直接决定了蒸馏知识与目标应用数据的表征一致性,影响了PrivateAI方法的性能。为了评估上述影响,我们进一步采用“通用对话”类数据集AIxBlock作为代理数据,从中采样了5 000条英文对话样本作为代理数据,上述数据集覆盖了用户的姓名、邮箱、卡号等多维度隐私敏感信息,用于蒸馏GPT-5的隐私识别能力。我们基于EMC-Smoke数据集及Transformer模型进行模型训练及实验评估,相关训练结果如表2所示。结果显示,当代理数据与目标应用数据在分布上存在显著差异,模型的训练效果受到明显影响,性能有所下降。然而,与完全不使用知识蒸馏的基线相比,即便代理数据集与目标数据存在异构性,引入代理知识仍能带来整体性能的提升。进一步验证了本文方法在跨领域场景下的有效性与价值。此外,我们也承认本文所提出的PrivateAI方法的

一个重要局限性:分布式训练模型性能优势依赖于代理数据集 P 的选取。虽然 PrivateAI 在同域场景中表现出显著优势,但在跨域应用中,其性能会随着代理数据与目标领域分布差异的增大而下降。然而在实际场景中收集与应用数据分布、领域接近的代理数据集通常是切实可行的,可通过以下 2 方面手段应对上述挑战:

(1)基于公开数据近似,通过收集与目标领域特征一致的开源语料或公开样本,构建与目标任务更贴近的代理数据集。

(2)基于语料合成,利用大模型生成与目标领域分布相似的合成语料作为代理数据,从而在保持隐私安全的前提下提升知识蒸馏的对齐性与有效性。

上述策略可在一定程度上降低 PrivateAI 对代理数据分布的依赖,提升其在跨领域任务中的适用性。

表 2 代理数据集分布对 PrivateAI 性能影响

Table 2 Impact of proxy dataset distribution on PrivateAI

类型	隐私识别精度
无代理数据	40.5
异构代理数据	68.3
同分布代理数据	78.1

4 结论

本文提出了一种新型隐私保护协同学习框架 PrivateAI,旨在保障在线生成式人工智能应用中的隐私安全。该框架能够协同利用分布式数据和基础模型中蕴含的丰富知识来训练隐私识别模型,同时保护训练数据中的隐私信息。通过在两个真实世界数据集上的大量实验,本文验证了 PrivateAI 框架所学习的隐私识别模型在隐私保护成功率方面相较于现有最优隐私保护机器学习方法提升 53.7 个百分点。PrivateAI 方法为在线人工智能应用用户隐私泄露问题提供了有效的解决方案,成功打破了隐私困境。此外,本研究还实现了对多源异质知识的高效提取与融合,突破了现有方法仅能利用编码于同质化载体知识的限制,大幅度提升了知识迁移的效率。总体而言,PrivateAI 不仅是一个具有实用价值的隐私保护框架,也为研究人员重新思考机器学习范式、探索分布式智能协同的可能路径提供了重要启发。此外,分布式训练模型性能优势依赖于代理数据集的选取。虽然 PrivateAI 在数据同分布场景中表现出显著优势,但在跨域应用中,其性能会随着代理数据与目标领域分布差异的增大而下降。然而,在实际场景中收集与应用数据分布接近的代理数据集通常是可行的,具体可以采用以下两种方案实现。一方面,基于公开数据近似,通过收集与目标领域特征一致的开源语料或公开样本,构建与目标任务更贴近的代理数据集。另一方面,基于语料合成,利用大模型生成与目标领域分布相似的合成语料作为代理数据,从而在保持隐私安全的前提下提升知识蒸馏的对齐性与有效性。通过采用上述策略,可以在一定程度上降低 PrivateAI 框架对代理数据分布的依赖性,从而有效提升其在跨领域任务中的泛化能力与适用性。

参考文献

[1] Kim M, Chen Chen, Wang Peng, et al. Detection of ovarian cancer via the spectral fingerprinting of quantum-defect-modified carbon nanotubes in serum by machine learning[J]. Nature Biomedical Engineering, 2022, 6(3): 267-275.

[2] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold[J]. Nature, 2021, 596(7873): 583-589.

[3] Reuters. ChatGPT sets record for fastest-growing user base (Analyst note) [EB/OL]. (2023-02-01) [2025-08-21]. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.

[4] Crompton H, Burke D. Artificial intelligence in higher education: The state of the field[J]. International Journal of Educational Technology in Higher Education, 2023, 20(1): 22.

[5] Gibney E. The scant science behind Cambridge Analytica's controversial marketing techniques[J]. Nature, 2018: 555: 286-287.

[6] Martin K D, Zimmermann J. Artificial intelligence and its implications for data privacy[J]. Current Opinion in Psychology, 2024, 58: 101829.

[7] Hartley J, Sanchez P P, Haider F, et al. Neural networks memorise personal information from one sample[J]. Scientific Reports, 2023, 13: 21366.

[8] Bloomberg. Samsung bans ChatGPT and other generative AI use by staff after leak[EB/OL]. (2023-05-02) [2025-08-21]. https://www.bloomberg.com/news/articles/2023-05-02/samsung-bans-chatgpt-and-other-generative-ai-use-by-staff-after-leak.

[9] Ren Xuebin, Yu Chia-Mu, Yu Weiren, et al. LoPub: High-dimensional crowdsourced data publication with local differential privacy[J]. IEEE Transactions on Information Forensics and Security, 2018, 13(9): 2151-2166.

[10] Norgeot B, Muenzen K, Peterson T A, et al. Protected Health Information filter (Philter): Accurately and securely de-identifying free-text clinical notes[J]. npj Digital Medicine, 2020, 3: 57.

[11] Warnat-Herresthal S, Schultze H, Shastry K L, et al. Swarm Learning for decentralized and confidential clinical

machine learning[J]. Nature, 2021, 594(7862): 265-270.

[12] Kaissis G A, Makowski M R, Rückert D, et al. Secure, privacy-preserving and federated machine learning in medical imaging[J]. Nature Machine Intelligence, 2020, 2(6): 305-311.

[13] Qi Tao, Wang Huili, Huang Yongfeng. Towards the robustness of differentially private federated learning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(18): 19911-19919.

[14] Mcmahan H B, Xu Zheng, Zhang Yanxiang. A hassle-free algorithm for strong differential privacy in federated learning systems[C]//Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track. Stroudsburg: ACL, 2024: 842-865.

[15] 康海燕, 王骁识. 基于数据特征相关性和自适应差分隐私的深度学习方法研究[J]. 电子学报, 2024, 52(6): 1963-1976.
Kang Haiyan, Wang Xiaoshi. Research on the deep learning method based on data feature relevance and adaptive differential privacy[J]. Acta Electronica Sinica, 2024, 52(6): 1963-1976. (in Chinese)

[16] Deußer T, Sparrenberg L, Berger A, et al. A survey on current trends and recent advances in text anonymization[C]//2025 IEEE 12th International Conference on Data Science and Advanced Analytics. Piscataway: IEEE, 2025: 11247969.

[17] Kovačević A, Bašaragin B, Milošević N, et al. De-identification of clinical free text using natural language processing: A systematic review of current approaches[J]. Artificial Intelligence in Medicine, 2024, 151: 102845.

[18] Sweeney L. Replacing personally-identifying information in medical records, the Scrub system[J]. Proceedings, 1996: 333-337.

[19] Robertson S, Zaragoza H. The probabilistic relevance framework: BM25 and beyond[J]. Foundations and Trends in Information Retrieval, 2009, 3(4): 333-389.

[20] Lample G, Ballesteros M, Subramanian S, et al. Neural architectures for named entity recognition[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: ACL, 2016: 260-270.

[21] Kenton J D, Chang Mingwei, Toutanova L K. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the NAACL, 2019: 4171-4186.

[22] Houman P, Montani I, Van Landschoot R, et al. spaCy: Industrial-strength natural language processing in Python[EB/OL]. (2024) [2025-09-09]. https://spacy.io/.

[23] Akbik A, Bergmann T, Blythe D, et al. FLAIR: An easy-to-use framework for state-of-the-art NLP[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations). Kerrville: Association for Computational Linguistics 2019: 54-59.

[24] Johnson A E W, Bulgarelli L, Pollard T J. Deidentification of free-text medical records using pre-trained bidirectional transformers[C]//Proceedings of the ACM Conference on Health, Inference, and Learning. New York: ACM, 2020: 214-221.

[25] Dou Yao, Krsek I, Naous T, et al. Reducing privacy risks in online self-disclosures with language models[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2024: 13732-13754.

[26] Mishra K, Pagare H, Sharma K. A hybrid rule-based NLP and machine learning approach for PII detection and anonymization in financial documents[J]. Scientific Reports, 2025, 15: 22729.

[27] Frikha A, Walha N, Nakka K K, et al. IncogniText: Privacy-enhancing conditional text anonymization via LLM-based private attribute randomization[C]//Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics, 2025: 2490-2501.

[28] Yang Tianyu, Zhu Xiaodan, Gurevych I. Robust utility-preserving text anonymization based on large language models[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2025: 28922-28941.

[29] Staab R, Vero M, Balunovic M, et al. Language models are advanced anonymizers[C/OL]//Proceedings of the ICLR, 2025, https://openreview.net/forum?id=82p8VHRsaK.

[30] McMahan B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data[C]//Proceedings of the AISTATS, 2017: 1273-1282.

[31] Reddi S J, Charles Z, Zaheer M, et al. Adaptive federated optimization[C/OL]//Proceedings of the ICLR, 2021, https://openreview.net/forum?id=LkFG3lB13U5.

[32] Li Tian, Sahu A K, Zaheer M, et al. Federated optimiza-

tion in heterogeneous networks[C]//Proceedings of the Third Conference on Machine Learning and Systems. Austin: MLSys, 2020, 2: 429-450.

[33] Dwork C, McSherry F, Nissim K, et al. Calibrating noise to sensitivity in private data analysis[M]//Theory of Cryptography. Berlin: Springer, 2006: 265-284.

[34] McMahan H B, Ramage D, Talwar K, et al. Learning differentially private recurrent language models[C/OL]//Proceedings of the ICLR, 2018, https://openreview. net/forum?id=BJ0hF1Z0b.

[35] Wei Kang, Li Jun, Ding Ming, et al. Federated learning with differential privacy: Algorithms and performance analysis[J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 3454-3469.

[36] Girgis A, Data D, Diggavi S, et al. Shuffled model of differential privacy in federated learning[C]//Proceedings of the AISTAS, 2021: 2521-2529.

[37] 赵登峰, 薛大暄, 赵素云, 等. 基于稀疏平滑自蒸馏的差分隐私深度学习方法[J]. 电子学报, 2025, 53(9): 3310-3318.
Zhao Dengfeng, Xue Daxuan, Zhao Suyun, et al. Differentially private with sparse and smooth self-distillation[J]. Acta Electronica Sinica, 2025, 53(9): 3310-3318. (in Chinese)

[38] Domingo-Ferrer J, Sánchez D, Blanco-Justicia A. The limits of differential privacy (and its misuse in data release and machine learning) [J]. Communications of the ACM, 2021, 64(7): 33-35.

[39] Kingma D P, Ba J. Adam: A method for stochastic optimization[C]//Proceedings of the ICLR, 2015.

[40] Uzuner O, Luo Yuan, Szolovits P. Evaluating the state-of-the-art in automatic de-identification[J]. Journal of the American Medical Informatics Association, 2007, 14(5): 550-563.

[41] Liu V, Musen M A, Chou T. Data breaches of protected health information in the United States[J]. Jama, 2015, 313(14): 1471.

[42] Stubbs A, Uzuner Ö. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus[J]. Journal of Biomedical Informatics, 2015, 58: S20-S29.

[43] Wu Xidong, Huang Feihu, Hu Zhengmian, et al. Faster adaptive federated learning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(9): 10379-10387.

[44] Sun Jingwei, Li Ang, Wang Binghui, et al. Soteria: Provable defense against privacy leakage in federated learning from representation perspective[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 9307-9315.

[45] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of the NeurIPS. Long Beach: Curran Associates Inc., 2017: 5998-6008.

[46] Liu Yinhan, Ott M, Goyal N, et al. RoBERTa: A robustly optimized BERT pretraining approach[PP/OL]. V1. arXiv (2019-07-26). https://doi.org/10.48550/arXiv.1907.11692.

[47] Carlini N, Chien S, Nasr M, et al. Membership inference attacks from first principles[C]//2022 IEEE Symposium on Security and Privacy. Piscataway: IEEE, 2022: 1897-1914.

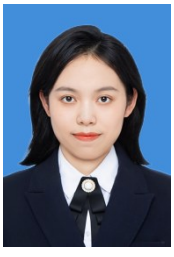
[48] Diao E M, Ding Jie, Tarokh V. SemiFL: Semi-supervised federated learning for unlabeled clients with alternate training[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022: 17871-17884.

作者简介

齐 涛 男, 1998 年 2 月出生于天津市。2024 年博士毕业于清华大学电子工程系。现为北京邮电大学研究员、博士生导师。主要研究方向大模型安全、AI 隐私计算等。
E-mail: taoqi@bupt.edu.cn

杨珮茹 女, 1999 年 8 月出生于江苏省连云港市。2021 年毕业于清华大学电子工程系本科。现为清华大学电子工程系博士研究生。主要研究方向为大模型及其安全优化。
E-mail: ypr21@mails.tsinghua.eud.cn

王慧丽 女, 1999 年 8 月出生于安徽省界首市。现为清华大学电子工程系博士研究生。主要研究方向为大模型及其安全。
E-mail: whl24@mails.tsinghua.edu.cn



王文丹 女, 2002 年 2 月出生于河南省濮阳市。现为北京邮电大学博士研究生。主要研究方向是检索增强生成、大模型安全。
E-mail: wendanwang@bupt.edu.cn

谭支鹏　男,1973年11月生于湖北省巴东县。2008年博士毕业于华中科技大学计算机系统结构专业。现为华中科技大学教授、博士生导师。主要研究方向大数据存储、AI存储与管理、移动存储等。中国电子学会会员编号:E190200129M。

E-mail: tanzhipeng@hust.edu.cn

黄永峰　男,1967年12月出生于湖北省赤壁市。2000年博士毕业于华中科技大学计算机系统结构专业。现为清华大学教授、博士生导师。主要研究方向数据安全、AI安全、信息隐藏等。

E-mail: yfhuang@tsinghua.edu.cn

王尚广　男,1982年2月出生于河南省许昌市。2011年毕业于北京邮电大学,获博士学位。现为北京邮电大学计算机学院教授、博士生导师。主要研究方向为服务计算、移动边缘计算与卫星计算。

E-mail: sgwang@bupt.edu.cn

徐红艳　女,1993年出生于河北省衡水市。2024年毕业于天津大学计算机应用技术专业。现为北京林业大学讲师。主要研究方向为自然语言处理、信息检索、大模型。

E-mail: hongyanxu@bjfu.edu.cn

罗传文　男,1991年出生于山东省菏泽市。2020年毕业于中国人民大学信息学院。现为北京林业大学副教授。主要研究方向为具身智能、边缘计算。中国电子学会会员编号:E190036595M。

E-mail: hongyanxu@bjfu.edu.cn