计算机研究与发展
Journal of Computer Research and Development
https://crad.ict.ac.cn
63(11): 2952−2962, 2026

本期推荐

面向检索增强生成敏感信息保护的抗测绘表征扰动

谭 学 郑 毅 张云若 陈怡然 陈 平 薛向阳
(复旦大学大数据研究院 上海 200433)
(xuetan23@m.fudan.edu.cn)

Anti-Mapping Representation Perturbation for RAG Sensitive Information Protection

Tan Xue, Zheng Yi, Zhang Yunruo, Chen Yiran, Chen Ping, and Xue Xiangyang
(*Institute of Big Data*, *Fudan University*, *Shanghai* 200433)

Abstract Retrieval-augmented generation (RAG) systems extend the capability of large language models by incorporating external knowledge databases, enabling more accurate and up-to-date responses. However, this integration also introduces a novel privacy vulnerability: mapping attacks (MAs), which aim to infer whether a specific private fragment is indexed in the retrieval database and to uncover its retrieval behavior. Such attacks exploit both embedding-level signals and retrieval dynamics, posing a serious threat to sensitive data protection in RAG pipelines. Despite the growing awareness of privacy risks in RAG systems, there is currently no dedicated defense mechanism specifically designed to mitigate mapping attacks. We introduce AMRP-SIP, a dual-randomization framework that concurrently protects both embeddings of the documents and retrieval traces, while preserving state-of-the-art utility. AMRP-SIP comprises three lightweight stages. First, a random orthogonal projection compresses each query and document into a low-dimensional latent space, hiding raw embeddings and reducing downstream noise. Second, adaptive differential privacy injects cluster-adaptive Gaussian noise, ensuring (ε, δ) fragment-level privacy. Third, a score-dropout layer introduces randomness by perturbing similarity scores with noise and probabilistically dropping a portion of the retrieved documents with probability p, thereby obfuscating the retrieval trajectory. Experiments on Wiki-40B, PubMed-256, and IP-DataSet demonstrate that AMRP-SIP reduces the AUC of membership inference attack (MIA) from 0.75 to 0.27, while maintaining competitive performance in downstream RAG tasks.

Key words retrieval-augmented generation (RAG); mapping attack (MA); differential privacy; membership inference attack (MIA); random orthogonal projection

摘 要 检索增强生成(retrieval-augmented generation, RAG)系统通过集成外部数据库扩展了语言模型的能力。然而,这种增强方式引入了一种新型隐私漏洞——测绘攻击(mapping attack, MA),该类攻击旨在推断某一特定的私有数据片段是否被收录于检索数据库中,并进一步揭示其在检索过程中的行为特征。这类攻击同时利用嵌入层面的信号和检索过程中的动态信息,对RAG流程中的敏感数据保护构成了严重威胁。尽管人们对RAG系统中的隐私风险关注不断提升,但目前尚不存在专门针对测绘攻击设计的防御机制。AMRP-SIP通过双重随机化机制同时保护文档嵌入表示与检索轨迹,并在隐私保护与检索效

收稿日期: 2025-09-30; **修回日期**: 2026-06-09
基金项目: 国家重点研发计划项目(2023YFB3107404); 宁波市“科创甬江2035”重点研发计划项目(2025Z054)
This work was supported by the National Key Research and Development Program of China (2023YFB3107404) and the Key Research and Development Program of Ningbo's "Science and Technology Innovation Yongjiang 2035" (2025Z054).
通信作者: 陈平(pchen@fudan.edu.cn)

用之间实现有效平衡。AMRP-SIP 包含 3 个轻量级阶段：首先，随机正交投影将查询和文档映射至低维潜在空间，以混淆原始嵌入表示，并降低后续噪声注入对检索效用的影响；其次，自适应差分隐私机制根据局部语义簇敏感度注入高斯噪声，确保实现 (ε, δ) 级别的文档片段隐私保护；最后，扰动丢弃阶段对相似度分数施加噪声扰动，并以概率 p 随机丢弃部分检索文档，从而模糊检索轨迹。在 Wiki-40B, PubMed-256, IP-DataSet 上的实验表明，AMRP-SIP 将成员推断攻击（membership inference attack, MIA）的 AUC 值从 0.75 降至 0.27。该框架在保护敏感信息的同时，维持了与现有方法相当的检索性能，为 RAG 系统中的测绘攻击防御提供了一种有效方案。

关键词 检索增强生成；测绘攻击；差分隐私；成员推断攻击；随机正交投影

中图法分类号 TP311.13；TP309

DOI: 10.7544/issn1000-1239.202550713 **CSTR:** 32373.14.issn1000-1239.202550713

大语言模型（large language model, LLM）展现出卓越的推理能力和流畅性，但其参数化记忆是静态的，在训练期间未接触过的知识始终无法被获取[1]。检索增强生成（retrieval-augmented generation, RAG）通过为每个用户输入整合从外部向量数据库检索的前 k 个文本段落[2-6]，解决了这一局限。虽然 RAG 实现了实时知识更新，却也引入了新的隐私风险[7]。测绘攻击（mapping attack, MA）允许攻击者通过精心构造的查询探测系统观察哪些文本段落被检索并返回。这类攻击不仅包含传统的成员推断攻击（membership inference attack, MIA）[8]，还会揭示文本片段在检索过程中的排名情况，对商业机密、医疗记录或受版权保护的文本构成重大风险。

近期的研究已经引入了专为 RAG 系统设计的 MIA。例如，S^2MIA 通过目标文档与生成回复之间的语义相似性来推断成员关系[9]，而 MBA 则通过遮蔽文档中的词语并评估 RAG 系统的预测准确性来检测成员关系[10]。现有的防御方法，如通过添加对抗性噪声的 MemGuard[11]，以及旨在减轻 RAG 特定风险的基于提示词的方法[12]，往往会损害系统的效用。这些局限性表明，亟需一种统一的防御机制，同时保护检索文档的嵌入表示与检索轨迹，以应对 MA 和 MIA 带来的联合隐私威胁。

在本文中，我们提出了 AMRP-SIP，一个旨在保护文档嵌入和检索轨迹的双重随机化防御框架，并在此基础上保持 RAG 系统的效用。AMRP-SIP 通过 3 个轻量级阶段实现这一目标。首先，随机正交投影（random orthogonal projection）将查询和文档映射到低维潜在空间，在保留主要语义关系的同时混淆原始嵌入表示，并降低后续噪声注入带来的效用损失。其次，自适应差分隐私（adaptive differential privacy）向文档嵌入中引入簇自适应的高斯噪声，在文档片段级别提供 (ε, δ)-隐私。这确保了敏感信息得到保护的同时维持了检索系统的排名准确性。最后，扰动检索文本的相似度分数，并按照固定概率抑制排名靠前的候选者，将攻击者在检索轨迹上的优势限制在 $1-p$ 以内，使其更难利用检索到的文档。

AMRP-SIP 引入的计算开销极小，在客户端仅需 1 次矩阵乘法。对文本嵌入的扰动可在离线阶段预计算，从而确保该防御机制在实际部署中具备良好的轻量化特性与可扩展性。我们在 3 个基准数据集（Wiki-40B, PubMed, IP-DataSet）上评估了 AMRP-SIP 的性能。实验结果表明，AMRP-SIP 显著降低了 MIA 的影响，将 AUC 从 0.75 降至 0.27，而 Recall@10 指标仅下降 2.1%，且第 95 百分位的延迟仅增加 15 ms。这些发现证实了 AMRP-SIP 在不牺牲系统效用的前提下，能有效防御 MA，具有很高的实用性。

总而言之，本文的贡献概括如下：

1）引入了 AMRP-SIP，一个新颖的双重随机化防御框架，可同时保护嵌入表示和检索轨迹。

2）我们提供了严谨的隐私分析，证明了 AMRP-SIP 为文档的嵌入表示实现了片段级别的 (ε, δ) 差分隐私保护，并将攻击者在检索轨迹上的优势在概率上界定为 $1-p$，从而为 RAG 系统中文档的嵌入表示和检索轨迹提供了端到端的隐私保障。

3）我们在多个基准数据集（Wiki-40B, PubMed, IP-DataSet）上进行了广泛的实验，结果表明 AMRP-SIP 在显著降低 MIA 有效性的同时，保持了较高的检索准确率和极低的延迟增量。

1 相关工作

1.1 大语言模型中的 MIA

近年来，大语言模型在广泛的自然语言处理任

务中表现出显著的能力,但同时引发了隐私方面的担忧,尤其是 MIA 的风险。Shokri 等人[8] 在深度学习环境下提出 MIA,揭示了即使是相对小规模的模型,也可能因对特定样本的过拟合而泄露训练数据。此后,多项研究表明,生成式语言模型相较于判别式模型更容易记忆并重现训练数据。例如,Carlini 等人[13] 发现 GPT-2 在未经防护的情况下,可能暴露大语言模型中的严重隐私风险。进而,后续研究利用多种方法定量分析了这类风险:Duan 等人[14] 发现,传统的基于影子模型的 MIA 攻击在 GPT-J, OPT 等模型上的效果仅略高于随机猜测,而 Panda 等人[15] 引入的更强探针“canary”能使 AUC 在标签可得的环境下超过 0.7。其他研究方向还探讨了微调或蒸馏模型[16]、视觉-语言对应模型[17] 以及基于差分隐私的防御机制[18]。综合来看,基础大语言模型的成员推断风险在严格防御条件下可能低于计算机视觉或分类模型,但在针对性查询构造下仍存在潜在脆弱性。

1.2 RAG 中的 MIA

RAG 系统结合了外部检索模块与大语言模型,并以外部文档作为上下文生成答案。然而,这类系统在隐私上存在新的挑战:尽管 RAG 能够提升生成答案的质量与事实依据性,但它同时也引入了新的 MIA 风险。Li 等人[9] 提出了 S^2MIA 方法,该方法利用候选文档与大语言模型回复之间的语义相似性来推断成员身份。Anderson 等人[12] 证明,精心设计的提示词可以在黑箱与灰箱场景中实现高 AUC 攻击效果,并提出在提示词中插入重新表述的指令,作为一种主要的防御手段。为了提高攻击的针对性,Liu 等人[10] 引入了 MBA 攻击,在这种方法中,目标文档中的部分 tokens 会被掩盖,如果模型能够完整补全这些位置,则说明它可能已经检索到了目标文档。Wang 等人[19] 进一步基于语义相似性进行了改进,结合难度感知似然比测试,将 TPR 在相同 FPR 条件下提升了 2 倍。最近,Naseh 等人[20] 提出了一种隐蔽的询问式攻击,仅需不超过 30 个自然语言“谜题”询问即可高效判定成员的身份。

1.3 针对大语言模型与 RAG 系统 MIA 的防御策略

最早针对大语言模型的 MIA 防御方法多将模型视作黑盒,并通过加入随机扰动降低攻击精度。MemGuard 方法通过向输出概率向量中加入精心设计的对抗性噪声,显著降低了攻击的 AUC,同时在功能性指标上仅产生轻微影响[11]。在 RAG 场景中,有一种轻量级但颇具实用性的变体,即在提示模板中插入“拒绝直接作答”的指令,从而用改写的句子替代原始回答[12]。更为系统化的方案尝试将差分隐私微调(differential privacy fine-tuning)应用于大语言模型,同时对参数进行裁剪与扰动[21]。近期的研究还将防护重点转向 RAG 系统的内部检索阶段以及通过检索轨迹暴露的敏感信息。例如,MIST 方法对编码器施加成员不变子空间的约束,并选择性地放松此约束以应对非易受攻击样本[2],在仅造成轻微准确率下降的情况下优于多数现有防护方法。在生成阶段,DPVoterRAG 对检索结果进行多数投票汇总,提供从检索到最终回答的端到端防护[22]。这些研究表明,能够同时防护向量嵌入与检索轨迹的联合通道防护机制,往往在隐私与可用性之间提供了最佳平衡性,尤其适用于大语言模型与 RAG 系统。

2 预备知识

2.1 RAG

RAG[23] 通过将外部知识检索与生成模型相结合,代表了自然语言处理领域的一项重大进展。与仅依赖于内部参数化记忆的传统模型不同,RAG 通过在推理时从大型外部知识库中检索相关文档来增强模型的能力。这使得模型能够生成更准确、上下文更丰富、知识更密集的响应。鉴于其灵活性,RAG 已在业界被广泛应用,例如在 WikiChat[18], FinGPT[24], ChatRTX[25] 等系统中;同时也在专业领域得到应用,如医疗诊断[26] 以及代码和邮件补全[27]。一个典型的 RAG 系统包含 2 个主要组件:一个是检索器(retriever),它根据查询选择最相关的文档;另一个是生成器(generator),它基于检索到的内容合成最终的响应。

2.2 差分隐私

差分隐私(differential privacy, DP)是一种保护隐私的框架[28],它保证了在数据集中包含或排除任意单个数据点时,计算结果不会发生显著变化。这是通过向输出中添加经过精确校准的噪声来实现的,从而使攻击者即使在完全了解其余数据集的情况下,也难以确定某个特定个体的数据是否被使用。形式上,对于任何相差 1 个元素的数据 D 和 D',以及对于输出空间的所有可测子集 S,一个随机化机制 $\mathcal{M}$ 若满足 ε-DP,则有

$$Pr[\mathcal{M}(D)\in S]\leqslant e^{\varepsilon}\times Pr[\mathcal{M}(D')\in S], \quad (1)$$

其中 ε 是隐私预算,用于控制添加噪声的量。较小的 ε 值以牺牲效用为代价,提供了更强的隐私保障。高斯机制是一种广泛用于实现差分隐私的方法,它通过向查询结果中添加噪声来保护隐私。给定一个

应用于数据集 D 的查询函数 Q，从均值为 0、方差根据 Q 的敏感度校准的高斯分布中采样噪声并添加到结果中。这确保了添加的噪声足以掩盖任何单个数据点的贡献，同时保持了输出的整体效用。差分隐私已被广泛应用于数据分析、机器学习和数据库查询系统，尤其是在医疗和金融等敏感领域，其重要性尤为突出。应用差分隐私的一个核心挑战在于平衡隐私与效用：更强的隐私保证（即更小的 ε）需要添加更多的噪声，这可能会降低结果的准确性和可用性。在机器学习中，差分隐私被用于训练带有噪声梯度的模型，确保模型不会记忆特定的数据点，从而防范成员推断攻击[29]。差分隐私在训练深度学习模型中的应用已在多种场景中得到探索，从图像分类到自然语言处理，它有助于降低模型对敏感数据过拟合的风险[30]。

3 前期准备

3.1 威胁模型

在本节中，我们为 AMRP-SIP 框架定义了威胁模型。其目标是刻画潜在攻击者的特征、能力以及他们试图通过利用 RAG 系统的漏洞所要达成的目的。这为评估我们的防御机制提供的安全保障奠定了基础。

攻击者旨在实现以下目标：

1）泄露敏感数据。攻击者的主要目标是从 RAG 系统中提取敏感信息，例如确定特定的私有文档是否被包含在知识库中，或者从检索到的内容中恢复特定的数据点。

2）推断系统行为。攻击者可通过分析系统在不同查询下的响应，推断检索阶段是否涉及特定敏感片段，并进一步分析查询、检索结果与生成输出之间的关联模式。通过仔细检查模型的输出，他们可以发现可能仍属机密的私有数据的残留痕迹，并利用这些信息来策划更有效的后续攻击。

针对 RAG 系统的攻击者可能拥有一系列能力，以利用其检索和生成组件中的漏洞。在我们的威胁模型中，我们假设攻击者具备以下能力：

1）访问检索过程。攻击者可能了解目标系统所采用的检索算法或使用相似的开源检索器进行模拟，并可通过黑盒查询观察系统输出，从而间接推断查询与文档之间的映射关系。这使他们能够观察和分析查询-文档的映射关系，这些关系可能揭示敏感的关联，或作为进一步攻击的基础。

2）观察系统输出。攻击者可以监控系统的响应，包括基于检索文档生成的输出。通过分析这些响应以及系统在不同查询下的行为，攻击者可以推断底层数据的属性，从而可能识别敏感信息，甚至推断出哪些数据点被包含在知识库中。

3.2 防御目标

AMRP-SIP 框架旨在通过降低 MA 和 MIA 的风险来加强 RAG 系统的隐私保护，同时保持系统性能。AMRP-SIP 主要从嵌入表示与检索轨迹 2 个层面提供保护，以降低攻击者推断特定数据片段是否被检索并参与生成的风险。通过融合差分隐私和噪声注入技术，该框架破坏了查询与敏感内容之间的关联，使得数据关系难以被追踪。此外，AMRP-SIP 在设计上引入了最小的性能开销，同时保持了高检索准确率和响应质量。这种在隐私保障与实用效率之间的平衡，使得 AMRP-SIP 成为一个可扩展且高效的防御方案。

4 AMRP-SIP

AMRP-SIP 是面向 RAG 系统设计的隐私保护框

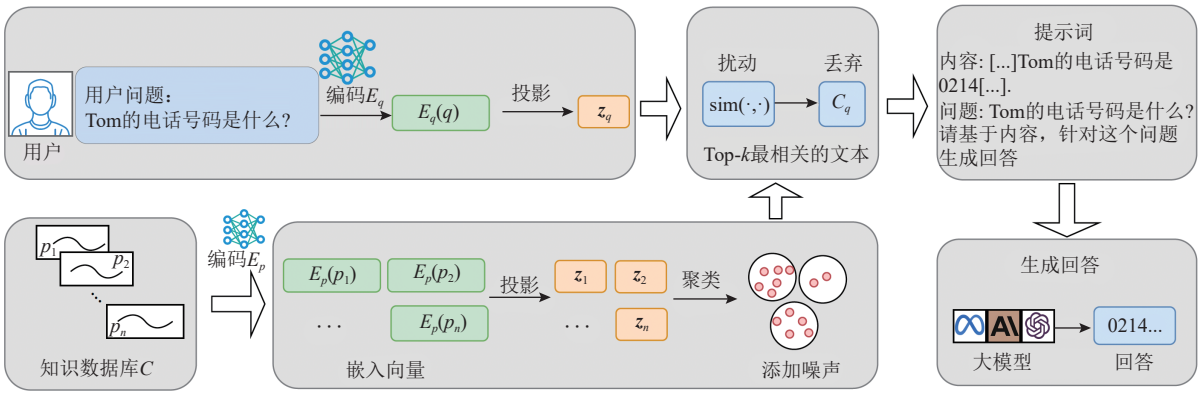


Fig. 1 The workflow of AMRP-SIP

图 1 AMRP-SIP 的工作流程

架, 旨在缓解 MIA 和 MA 带来的隐私风险, 同时兼顾系统的运行效率与实用性, 其详细流程如图 1 所示。该解决方案引入了随机正交投影、自适应差分隐私(adaptive-DP)和扰动丢弃等技术。AMRP-SIP 系统流程的算法如算法 1 所示。

算法 1. AMRP-SIP 系统流程。

输入: 查询 q, 知识库 C, 隐私参数 (ε, δ), 丢弃概率 p_{drop}, 噪声标准差 σ;

输出: 生成的回复 y。

① 查询编码与投影: 编码查询 $\boldsymbol{e}_q = f(q)$, 投影查询 $\boldsymbol{z}_q = \boldsymbol{R}\cdot\boldsymbol{e}_q$;

② 上传保护向量: 投影查询 $\boldsymbol{z}_q = \boldsymbol{R}\cdot\boldsymbol{e}_q$, 将 $\boldsymbol{z}_q$ 发送至服务器(本地保留 $\boldsymbol{e}_q$), 服务器离线预处理;

③ 文档编码与投影: 文档编码 $\boldsymbol{e}_d = f(d_i)$, 文档投影 $\boldsymbol{z}_d = \boldsymbol{R}\cdot\boldsymbol{e}_d$;

④ 自适应差分隐私: 添加噪声 $\tilde{\boldsymbol{z}}_d = \boldsymbol{z}_d + \mathcal{N}(0,\sigma_j^2 I_m)$, 其中 $\sigma_j = \dfrac{\Delta_j\sqrt{2\ln(1.25/\delta)}}{\varepsilon}$;

⑤ 构建向量索引: 使用构建索引(HNSW/IVF-PQ), 服务器在线查询处理;

⑥ 检索排名最高的候选项: 使用 ANN 搜索 $S_{\text{raw}} = ANN(\boldsymbol{z}_q)$;

⑦ 应用扰动丢弃: 添加噪声 $noisy_scores = scores + \mathcal{N}(0,\sigma^2)$, 应用丢弃以概率 $1-p_{\text{drop}}$ 保留候选, 其余丢弃, 若 $keep_i = 0$, $noisy_scores_i \leftarrow -1\times 109$;

⑧ 选择前 TOP-k 文档: 从 S_{raw} 中选取排名前 k 的文档 S_{top};

⑨ 构建提示(*Prompt*): 拼接查询与文档内容 $Prompt = concat\,(q, S_{\text{top}})$;

⑩ 生成回复: 使用大语言模型生成回复 $y = LLM\,(Prompt)$, 返回 y。

4.1 查询编码与上传保护

在客户端, 用户的查询 q 首先通过一个轻量级的句子嵌入模型(如 Contriever[31] 或 DPR-mul[32])编码, 生成一个 768 维的查询向量 $\boldsymbol{e}_q$。在此阶段, 该查询向量仅为一个可用于后续检索的语义向量, 但尚未经过隐私保护, 因此可能泄漏用户查询的语义信息, 甚至被攻击者近似重构原始查询。为防止这种情况, 查询向量 $\boldsymbol{e}_q$ 经过一个随机正交投影矩阵 $\boldsymbol{R}$ 的处理, 生成一个低维向量 $\boldsymbol{z}_q = \boldsymbol{R}\cdot\boldsymbol{e}_q$, 其中随机正交投影矩阵 $\boldsymbol{R}\in\mathbb{R}^{m\times d}$ 将查询投影到一个低维空间。该过程利用了 JL(Johnson-Lindenstrauss)定理[14], 确保相似度失真保持在 5% 以下, 同时显著增加逆向恢复原始查询向量的难度, 从而增强了隐私保护。最终, 经过保护的查询向量 $\boldsymbol{z}_q$ 被上传至服务器, 而原始查询向量 $\boldsymbol{e}_q$ 仍保留在客户端, 防止了原始查询数据的暴露。此外, 随机正交投影通过压缩高维嵌入降低传输开销, 从而提升客户端的整体效率(对应算法 1 中的步骤①~③)。

4.2 文档编码与差分隐私处理

在服务器端, 所有文档 d 首先使用相同的嵌入模型(例如 Contriever[31])编码为文档向量 $\boldsymbol{e}_d$。这些向量与查询嵌入位于同一语义空间, 并为后续检索做好准备。为了提供初步的隐私保护, 每个文档向量 $\boldsymbol{e}_d$ 都通过相同的随机正交矩阵 $\boldsymbol{R}$ 投影到一个低维空间, 得到转换后的表示 $\boldsymbol{z}_d = \boldsymbol{R}\cdot\boldsymbol{e}_d$。该投影在降低维度和混淆原始文档嵌入的同时, 保留了查询与文档之间的相对语义关系。为进一步增强隐私, 我们应用了自适应差分隐私。首先, 使用 k-NN 算法[33] 将文档向量 $\boldsymbol{z}_d$ 划分到不同的语义簇 C_j 中。对于每个簇, 计算其局部敏感度 $\Delta_j = \max\limits_{\boldsymbol{u},\boldsymbol{v}\in C_j}\|\boldsymbol{u}-\boldsymbol{v}\|_2$, 其中 $\boldsymbol{u}$ 和 $\boldsymbol{v}$ 是该簇内的任意 2 个文档向量。然后, 将如下的高斯噪声添加到每个文档向量 $\boldsymbol{z}_d$ 中:

$$\sigma_j = \frac{\Delta_j\sqrt{2\ln(1.25/\delta)}}{\varepsilon},\ \tilde{\boldsymbol{z}}_d = \boldsymbol{z}_d + \mathcal{N}\left(0,\sigma_j^2 I_m\right)。\quad (2)$$

这一步骤确保了文档嵌入向量满足差分隐私要求, 其添加的噪声根据局部敏感度进行了校准。因此, 该机制在最小化隐私泄露的同时, 保持了检索过程中排序的完整性(对应算法 1 中的步骤④⑤)。

4.3 向量数据库的构建与检索

加噪后的文档向量 $\tilde{\boldsymbol{z}}_d$ 被用于构建基于 HNSW 或 IVF-PQ 的向量索引[4,34], 以实现高效检索。在此阶段, 由于已在噪声注入过程中对文档向量进行扰动, 因此检索构建过程中无需再额外引入噪声。在查询阶段, 客户端将查询向量 $\boldsymbol{z}_q$ 发送至服务器, 由服务器执行近似最近邻(ANN)搜索[34], 以检索排名前 $k+k_{\text{ex}}$ 的候选文档。候选文档数量故意设置为多于 k 个, 以确保在应用扰动丢弃(score-dropout)后, 仍有足够数量的相关文档可用于生成回复(见算法 1 的步骤⑥⑦)。

4.4 扰动丢弃机制与回复生成

在候选检索阶段, 首先对查询和候选文档间的相似度分数添加均值为 0、方差为 $\sigma = 0.05$ 的高斯噪声。随后, 以概率 p_{drop} 随机移除一部分候选文档。该机制有效模糊了检索轨迹, 降低了攻击者基于分数分布推断敏感文档存在的可能性。剩余的候选文档根据扰动后的分数进行排序, 选取前 k 个文档。这些文档与原始查询拼接后构造成 Prompt, 并传递给大语言模型以生成上下文感知的安全回复。此过程在

保证上下文相关性的同时，提供了对查询及返回内容的多重隐私保护（对应算法 1 的步骤⑧~⑩）。

5 AMRP-SIP 的隐私性分析

AMRP-SIP 框架集成了多种隐私保护技术，包括随机正交投影、差分隐私及扰动丢弃，以确保敏感的用户查询和文档信息在处理过程中始终受到保护。

1）随机正交投影。系统首先对查询向量$\boldsymbol{e}_q$施加随机正交变换，生成维度较低的投影向量$\boldsymbol{z}_q = \boldsymbol{R}\cdot\boldsymbol{e}_q$。该变换使得攻击者在未知测绘矩阵$\boldsymbol{R}$的情况下，难以恢复原始查询。该特性符合 JL 引理[35]，即在随机线性投影下，向量对之间的欧氏距离能够以高概率近似保留。对于任意 2 个向量$\boldsymbol{u},\boldsymbol{v}\in\mathbb{R}^d$，一个合适的投影矩阵$\boldsymbol{R}\in\mathbb{R}^{m\times d}$能确保：

$$\|\boldsymbol{R}\cdot\boldsymbol{u}-\boldsymbol{R}\cdot\boldsymbol{v}\|_2 \leqslant \epsilon\|\boldsymbol{u}-\boldsymbol{v}\|_2\,, \tag{3}$$

其中ϵ是一个较小的失真界限，其值取决于目标维度m以及被投影向量的数量。

在服务器端，通过向文档向量$\boldsymbol{z}_d$中添加高斯噪声来实现自适应差分隐私，以提供形式化的差分隐私保障。噪声的量级根据文档空间中每个语义簇C_j的局部敏感度Δ_j进行校准。具体而言，噪声尺度σ_j定义为：

$$\sigma_j = \frac{\Delta_j\sqrt{2\ln(1.25/\delta)}}{\varepsilon},\ \tilde{\boldsymbol{z}}_d = \boldsymbol{z}_d + \mathcal{N}\left(0,\sigma_j^2 I_m\right)。 \tag{4}$$

为了严格证明上述机制满足差分隐私约束，我们给出定理 1。

定理 1. 差分隐私保证。对于任意相邻数据集D和D'（即相差一个文档片段），以及给定的隐私参数$\varepsilon>0$和$\delta\in(0,1)$，若对文档向量添加高斯噪声$\mathcal{N}\left(\sigma_j^2 I_m\right)$，且噪声尺度$\sigma_j$满足$\sigma_j \geqslant \frac{\Delta_j\sqrt{2\ln(1.25/\delta)}}{\varepsilon}$，则该机制满足$(\epsilon,\delta)$-差分隐私。

证明. 根据 Dwork 等人[28]提出的高斯机制（Gaussian mechanism）理论，对于L_2-敏感度为Δ_f的查询函数f，当添加噪声$\sigma \geqslant \frac{\Delta_j\sqrt{2\ln(1.25/\delta)}}{\varepsilon}$时，机制满足$(\varepsilon,\delta)$-DP。在本方案中，我们针对每个语义簇$C_j$计算局部敏感度$\Delta_j = \max\limits_{\boldsymbol{u},\boldsymbol{v}\in C_j}\|\boldsymbol{u}-\boldsymbol{v}\|_2$。通过式（4）自适应地将噪声校准至$\Delta_j$，我们确保了在簇级别的扰动满足上述高斯机制的条件。因此，该方法在文档片段级别实现了(ε,δ)-差分隐私，从而防止攻击者仅凭含噪表示$\tilde{\boldsymbol{z}}_d$就能推断特定文档的存在。 证毕.

2）扰动丢弃。在初步检索后，对候选文档的相似度分数加入小幅噪声，并以概率p_{drop}随机丢弃部分候选。该过程进一步隐藏检索轨迹，降低了敏感文档暴露风险。攻击者在连续k次选择操作中猜中某一文档的概率最大值受式（5）约束。

$$Adv_{\text{MIA}} \leqslant (1-p_{\text{drop}})^k。 \tag{5}$$

这种概率抑制机制确保了即使对手能够观察多组查询-输出结果对，其准确推断出某一特定文档是否被包含的可能性仍然很低。

AMRP-SIP 在防御 MIA 与 MA 方面提供了强有力的保护，同时保持了高效的检索与模型性能。虽然扰动丢弃和差分隐私等机制在引入隐私保护的同时会带来一定的准确率下降，但当丢弃率较低或隐私预算设置不过于严格时，其整体性能影响依然有限。AMRP-SIP 在隐私性与可用性之间实现了实用且灵活的平衡，非常适合在隐私敏感型应用中部署。

6 实　　验

6.1 实验设置

在本节中，我们介绍实验的设置。

6.1.1 数据集

我们构建了涵盖多种场景的 3 类检索语料库，包括互联网、开放领域和科学语料。1）恶意 IP 数据集（IP-DataSet）。由 ZoomEye[36]平台采集的 IP 地址构成，涵盖地理位置、ISP 及相关攻击记录，可用于评估检索系统抵御恶意查询的能力。2）Wiki-40B。基于英文维基百科[37]快照清洗生成，包含 530 万文段，并保留文内结构。3）PubMed-256。由 PubMed 2025 基于生物医学文献生成，共含 215 万个生物医学文段[38]。

所有语料均按固定长度截断。Wiki-40B 每段最多保留 512 个 token，PubMed-256 和 IP-DataSet 每段最多保留 256 个 token。截断后的文段使用无监督的 Contriever 编码器编码为 768 维向量，并通过 FAISS HNSW 图[34]（$M=2$, *efSearch* = 128）进行对数时间近似最近邻检索。我们在评估前随机抽取 10% 的文段作为被检索的成员集合，并在此集合上运行整个 AMRP-SIP 流程。

6.1.2 攻击者

为了评估 AMRP-SIP，我们复现了针对 RAG 系统的 MIA，包括：1）S^{2}MIA[9]。利用候选文档与 RAG 生成回答之间的语义相似性进行成员关系推断。2）掩码基推断（mask-based MIA, MBA）[10]。随机掩盖文段中 15% 的 token，并基于模型的精确匹配概率来进行判断。3）DC-MIA（RAG-leaks）[19]。对查询-答案对执行成员可能性打分，进行阈值判定。4）询问攻击

(interrogation attack, IA)[20]。手工构建多轮自然语言“细粒度”问题,其中只有在检索到的目标文段存在时才能准确作答。我们在复现中严格按照原文公开的实现与超参数设置,包括选取前5%扰动最高的检索查询、限制每篇文档的API请求量最多30次,并在1 000条查询的平衡集合上执行黑盒测试。

6.1.3 对比方案

在评估中,我们比较了4类防御策略。

1)Template-Block。在RAG提示中追加“禁止直接回答”指令,不引入额外推理[12]。

2)MemGuard-R。依据MemGuard原理精心添加噪声来抑制输出概率向量中的差异,使成员与非成员样本难以区分[11]。

3)MIST[2](membership-invariant subspace training)。通过对抗性子空间正则化编解码器权重,防止过拟合而导致的易受成员推断攻击样本,同时保持模型性能。

4)AMRP-SIP(本文方法)。在随机正交投影、自适应差分隐私与扰动丢弃的多重保护下抵御查询与检索轨迹的隐私泄露。

6.1.4 评估指标

1)隐私性。使用ROC曲线下面积(AUC)量化,衡量真阳性率与假阳性率的整体区分能力。

2)检索质量。以Recall@K衡量,即前K条检索结果中包含目标文段的比例。

3)效率。采用的是第95百分位延迟(P95 latency),它表示有95%的请求可以在这个时间内完成。

6.1.5 实现细节

所有查询和文档首先使用Contriever模型嵌入为768维的向量。随后,每个向量都通过一个矩阵$\boldsymbol{R}\in\mathbb{R}^{307\times768}$进行投影,该矩阵满足JL失真界限。接着,通过注入簇自适应的高斯噪声对向量进行扰动,以实现片段级别的度量差分隐私(ε,δ)保障。经过处理后的带噪嵌入被离线存储在一个FAISS HNSW索引中($M=32$, $efSearch=128$),该索引部署在双NVIDIA A100-80GB GPU上。在推理时,客户端执行1次$\boldsymbol{R}\cdot\boldsymbol{e}_q$的投影操作;服务器则检索最邻近的$k+k_{\text{ex}}=10$个邻居,并应用丢弃概率为$p=0.3$、噪声标准差为$\sigma\approx0.05$的扰动丢弃策略。我们报告了5次独立运行的平均性能,并使用威尔科克森符号秩检验(Wilcoxon signed-rank test)来评估统计显著性,其显著性水平阈值设为$p\leqslant0.05$。

6.2 AMRP-SIP的主要实验结果

在GPT-4o作为大语言模型的条件下,我们对4种成员推断攻击(S^{2}MIA, MBA, DC-MIA, IA)在3个检索语料(Wiki-40B, PubMed, IP-DataSet)上的防护效果进行了全面评测。如表1所示,AMRP-SIP展现出了超越现有防御手段的隐私保护能力。具体而言,无防护系统在面对各类攻击时表现脆弱,AUC高达0.58~0.84。这是因为攻击者能够轻易利用查询与检索结果之间的高度语义相关性进行推断。对比方案中,Template-Block仅在提示词层面进行阻断,未能触及底层的检索链路,因此防御效果微乎其微。MemGuard-R在RAG场景下效果受限,因为它主要通过扰动输出概率分布来防御,而无法掩盖RAG特有的在检索阶段暴露的文本相似度特征。相比之下,AMRP-SIP通过在向量空间和检索排序2个维度的同时引入随机性,显著将所有攻击的AUC降低至0.22~0.40。这种性能的提升主要归功于AMRP-SIP切断了攻击者依赖的“确定性映射”,即攻击者不再确信高相似度文档必然被检索,也无法通过微小的分数差异来逆向推导文档是否存在。

此外,表1的Recall@10数据显示,AMRP-SIP相比无防御基准仅下降了2.1个百分点(53.4%降至51.3%)。这表明基于局部敏感度校准的噪声并未显著破坏检索的准确性,查询仍能精准命中目标文本。结果证实,AMRP-SIP在将隐私攻击AUC从0.84显著降至0.40的同时,有效维持了解决实际问题的核

Table 1 AUC and Recall@10 for Four Membership-Inference Attacks on Three Datasets

表1 3种数据集上4种成员推断攻击的AUC与Recall@10

防御方法	AUC(Wiki-40B)				AUC(PubMed-256)				AUC(IP-DataSet)				Recall@10/%
	S^{2}MIA	MBA	DC-MIA	IA	S^{2}MIA	MBA	DC-MIA	IA	S^{2}MIA	MBA	DC-MIA	IA	
无防御	0.84	0.80	0.82	0.63	0.81	0.77	0.78	0.60	0.72	0.70	0.75	0.58	53.4
Template-Block	0.66	0.60	0.62	0.45	0.62	0.57	0.59	0.42	0.53	0.50	0.55	0.40	53.2
MemGuard-R	0.61	0.56	0.59	0.40	0.58	0.54	0.55	0.38	0.49	0.46	0.50	0.35	52.7
MIST	0.53	0.49	0.50	0.36	0.50	0.46	0.47	0.34	0.44	0.41	0.43	0.31	49.9
AMRP-SIP(本文)	**0.40**	**0.37**	**0.35**	**0.29**	**0.38**	**0.35**	**0.33**	**0.27**	**0.30**	**0.28**	**0.27**	**0.22**	**51.3**

注:AUC值越低且Recall@10值越高,则表现越优。黑体数值表示最优值。

心检索能力。

6.3 消融研究

为了深入探究 AMRP-SIP 的防御机理，表 2 消融实验揭示了 3 层防御逻辑的贡献。首先，仅部署随机正交投影（ROP）时，AUC 降至 0.60~0.68，这是由于投影将高维语义向量压缩至低维潜在空间，实现了对高维语义空间的几何混淆，有效破坏了攻击者基于欧氏距离的粗粒度测绘能力。其次，引入自适应差分隐私（DP）后，AUC 进一步降至 0.40~0.50，其核心在于利用局部敏感度校准噪声掩盖了文档的微观特征，使攻击者难以区分语义簇内的具体私有片段，从而有效抵御了针对特定数据点的细粒度成员推断。最后，结合扰动丢弃（Drop）将 AUC 降低至最低点（0.22~0.40）。该机制通过概率性移除高分候选者，直接阻断了检索排序的确定性逻辑，迫使现有的基于“高相似度必高排名”假设的阈值判定模型失效，从而提供了强力度的防护。

6.4 隐私与可用性权衡

本节深入探讨 AMRP-SIP 在不同隐私强度下的效能表现，这种权衡主要体现在差分隐私预算 ε 与扰动丢弃策略 2 个维度。表 3 展示了在完整 AMRP-SIP 防御策略下，当隐私预算 ε 从 0.5 逐步调整至 4.0 时，系统在隐私性（以 AUC 衡量，数值越低表示攻击越难成功）与可用性（以 Recall@10 衡量，数值越高表示检索效果越好）之间的变化关系。表 3 覆盖了 Wiki-40B, PubMed-256, IP-DataSet 这 3 个语义分布差异显著的数据集，并在每个数据集上独立重复 3 次实验以缓解随机性因素，获得更稳定的评估结果。实验结果表明，在所有数据集及 4 种攻击设置下，较小的 ε 值（即隐私预算收紧、噪声增大）能够一致性地降低攻击的 AUC，隐私性更好，而仅导致 Recall@10 极小幅度的下降。这一现象表明，AMRP-SIP 在不同语义

Table 2 AUC Values for Four Attacks on Three Datasets

表 2 3 种数据集上 4 种攻击的 AUC 值

方法变体	Wiki-40B				PubMed-256				IP-DataSet			
	S^2MIA	MBA	DC-MIA	IA	S^2MIA	MBA	DC-MIA	IA	S^2MIA	MBA	DC-MIA	IA
无防御	0.84	0.80	0.82	0.63	0.81	0.77	0.78	0.60	0.72	0.70	0.75	0.58
ROP	0.68	0.64	0.65	0.48	0.66	0.62	0.63	0.46	0.57	0.55	0.60	0.44
ROP+DP	0.50	0.47	0.46	0.36	0.48	0.45	0.44	0.34	0.42	0.40	0.40	0.31
ROP+DP+Drop	**0.40**	**0.37**	**0.35**	**0.29**	**0.38**	**0.35**	**0.33**	**0.27**	**0.30**	**0.28**	**0.27**	**0.22**

注：黑体数值表示最优值。

Table 3 Privacy–Utility Trade-Off for AMRP-SIP

表 3 AMRP-SIP 的隐私-效用权衡

数据集	ε	第 1 次测试					第 2 次测试					第 3 次测试				
		AUC				Recall@10/%	AUC				Recall@10/%	AUC				Recall@10/%
		S^2MIA	MBA	DC-MIA	IA		S^2MIA	MBA	DC-MIA	IA		S^2MIA	MBA	DC-MIA	IA	
Wiki-40B	0.5	0.32	0.30	0.28	0.24	49.6	0.35	0.32	0.30	0.25	48.9	0.26	0.24	0.23	0.19	47.8
	1.0	0.40	0.37	0.35	0.29	51.3	0.43	0.40	0.37	0.31	50.7	0.34	0.32	0.30	0.25	49.4
	2.0	0.48	0.45	0.42	0.36	52.2	0.52	0.48	0.46	0.39	51.8	0.41	0.39	0.37	0.31	50.5
	4.0	0.60	0.57	0.54	0.46	52.9	0.62	0.59	0.56	0.48	52.6	0.53	0.51	0.48	0.40	51.3
PubMed-256	0.5	0.30	0.28	0.27	0.23	48.8	0.33	0.31	0.29	0.25	48.1	0.24	0.23	0.22	0.18	46.9
	1.0	0.38	0.35	0.33	0.27	50.5	0.41	0.38	0.36	0.30	50.0	0.30	0.28	0.26	0.22	48.7
	2.0	0.46	0.44	0.41	0.34	51.4	0.49	0.46	0.44	0.37	51.2	0.38	0.36	0.34	0.28	49.8
	4.0	0.58	0.56	0.52	0.44	52.0	0.60	0.58	0.55	0.47	51.7	0.50	0.48	0.46	0.38	50.6
IP-DataSet	0.5	0.26	0.24	0.23	0.19	45.9	0.29	0.27	0.26	0.22	45.3	0.20	0.19	0.18	0.15	44.1
	1.0	0.30	0.28	0.27	0.22	47.6	0.33	0.31	0.30	0.25	47.0	0.24	0.23	0.22	0.18	45.9
	2.0	0.38	0.35	0.34	0.28	48.7	0.40	0.38	0.36	0.31	48.3	0.32	0.30	0.29	0.24	47.0
	4.0	0.48	0.46	0.44	0.36	0.36	0.50	0.48	0.46	0.39	49.1	0.41	0.40	0.38	0.32	47.8

注：每个区块列出 S^2MIA，MBA，DC-MIA，IA 的 AUC 值及 Recall@10 指标，随ε值的变化情况。各列分别对应 Wiki-40B，PubMed-256，IP-DataSet 数据集重复 3 次。

场景下均能够提供稳定且可调节的隐私-效用权衡能力, 在显著增强隐私保护的同时保持平稳的可用性退化。

此外, 针对“扰动丢弃”可能误删关键文档的风险, AMRP-SIP 采用了“预检索扩充”策略, 即基于 $k+k_{ex}$ 个候选构建缓冲池。关键文档凭借高相似度优势, 在扩充池中具有极高的留存率。这种检索的不确定性是防御测绘攻击的必要代价; 实验数据也证实了其合理性。在抑制 MIA 的 AUC 至 0.22 的强隐私保护下, Recall@10 相比基准仅下降 2.1 个百分点(见表 1)。这表明该机制有效平衡了极端场景下的丢弃风险与整体检索效用。对于准确性要求极高的特殊应用场景, 用户也可通过调低丢弃概率 p_{drop} 或增加扩充数量 k_{ex} 来进一步动态调整系统的鲁棒性。

6.5 检索效率评估

表 4 报告了在 Wiki-40B 检索语料下, 以典型工作负载 S^{2}MIA 攻击(针对整个检索过程)为例, 不同防御方法的端到端第 95 百分位延迟。Template-Block 几乎不引入额外延迟; MemGuard-R 由于需要在运行时添加噪声, 延迟增加约 4 ms; MIST-E 的延迟增加约 5 ms, 原因是使用了较大的编码器模型; AMRP-SIP 额外引入 1 次投影与扰动丢弃操作, 总延迟增加约 15 ms, 仍完全在 60 ms 的交互式延迟预算以内。

Table 4 P95 Latency for a Single S^{2}MIA Query on Each Dataset

表 4 单个 S^{2}MIA 查询每个数据集的 P95 延迟 ms

防御方法	Wiki-40B	PubMed-256	FinanceBench
无防御	41	43	38
Template-Block	41	43	38
MemGuard-R	44	46	42
MIST	46	48	44
AMRP-SIP (本文)	56	58	53

7 结论与未来工作

本文介绍了 AMRP-SIP, 一种面向 RAG 系统的隐私保护框架, 能够在不显著影响系统性能的情况下, 有效缓解 MIA 和 MA。AMRP-SIP 结合随机正交投影、指标感知差分隐私与扰动丢弃, 在提供强隐私保障的同时, 将对检索准确率的影响降至最低。实验结果表明, AMRP-SIP 能在显著降低多种攻击成功率的同时, 仅造成极小的召回率下降。AMRP-SIP 的模块化设计确保了其与多种检索和生成模型的兼容性, 并可无缝集成至现有的 RAG 流水线[39-40]。未来的工作方向包括: 持续优化各类防御机制; 探索更细颗粒度的隐私预算调节策略; 将该框架扩展到多模态的 RAG 任务中。

作者贡献声明: 谭学主导抗测绘表征扰动框架(AMRP-SIP)的设计与实现, 完成论文主要撰写工作; 郑毅负责部分代码实现; 张云若负责辅助框架设计; 陈怡然负责论文修改与审核; 陈平、薛向阳指导方案构建和论文修改。

参 考 文 献

[1] Zhu Jiachen. Research on attack and defense technologies of large language models for generated content security[D]. Chengdu: University of Electronic Science and Technology of China, 2025 (in Chinese)
(朱家辰. 面向生成内容安全的大语言模型攻防技术研究[D]. 成都: 电子科技大学, 2025)

[2] Li Jiacheng, Li Ninghui, Ribeiro B. MIST: Defending against membership inference attacks through membership-invariant subspace training[C]//Proc of the 33rd USENIX Security Symp. Berkeley, CA: USENIX Association, 2024: 2387−2404

[3] Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[C]//Proc of the 34th NeurIPS. Red Hook, NY: NeurIPS Foundation, 2020: 9459−9474

[4] Jgou H, Douze M, Schmid C. Product quantization for nearest neighbor search[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(1): 117−128

[5] Chen Donglei, Qin Wei. Research on key technologies of intelligent question answering system based on retrieval-augmented generation [J]. Office Informatization, 2024, 29(19): 82−86 (in Chinese)
(陈冬雷, 秦薇. 基于检索增强生成的智能问答系统关键技术研究[J]. 办公自动化, 2024, 29(19): 82−86)

[6] Bao Ganglin. Research on retrieval-augmented generation method of large language models for knowledge question answering[D]. Beijing: Beijing Jiaotong University, 2024 (in Chinese)
(包刚林. 面向知识问答的大语言模型检索增强生成方法研究[D]. 北京: 北京交通大学, 2024)

[7] Shi Jiayu, Chen Bo, Yu Li. Detection of stealthy FDI attacks based on Laplacian eigenmapping learning[J]. Acta Automatica Sinica, 2021, 47(10): 2494−2500 (in Chinese)
(石家宇, 陈博, 俞立. 基于拉普拉斯特征测绘学习的隐匿 FDI 攻击检测[J]. 自动化学报, 2021, 47(10): 2494−2500)

[8] Shokri R, Stronati M, Song Congzheng, et al. Membership inference attacks against machine learning models[C]//Proc of the 38th IEEE S&P. Piscataway, NJ: IEEE, 2017: 3−18

[9] Li Yuying, Liu Gaoyang, Wang Chen, et al. Generating is believing: Membership inference attacks against retrieval-augmented generation[C]//Proc of the 50th ICASSP 2025. Piscataway, NJ: IEEE, 2025: 1−5

[10] Liu Mingrui, Zhang Sixiao, Long Cheng. Mask-based membership inference attacks for retrieval-augmented generation[C]//Proc of the 34th ACM Web Conf 2025. New York: ACM, 2025: 2894−2907

[11] Jia Jinyuan, Salem A, Backes M, et al. MemGuard: Defending against black-box membership inference attacks via adversarial examples[C]//Proc of the 26th ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2019: 259−274

[12] Anderson M, Amit G, Goldsteen A. Is my data in your retrieval database? Membership inference attacks against retrieval-augmented generation[C]//Proc of the 11th Int Conf on Information Systems Security and Privacy (ICISSP 2025). Setúbal: SciTePress, 2025: 474–485

[13] Carlini N, Tramer F, Wallace E, et al. Extracting training data from large language models[C]//Proc of the 30th USENIX Security Symp. Berkeley, CA: USENIX Association, 2021: 2633−2650

[14] Duan M, Suri A, Mireshghallah N, et al. Do membership inference attacks work on large language models?[J]. arXiv preprint, arXiv: 2402.07841, 2024

[15] Panda A, Tang X, Nasr M, et al. Privacy auditing of large language models[J]. arXiv preprint, arXiv: 2503.06808, 2025

[16] He Yu, Liu Liu, Li Yiming, et al. Towards label-only membership inference attacks against pre-trained LLMs[C]//Proc of the 34th USENIX Security Symp. Berkeley, CA: USENIX Association, 2025: 1−18

[17] Li Zhan, Wu Yongtao, Chen Yihang, et al. Membership inference attacks against large vision-language models[C]//Proc of the 38th Advances in Neural Information Processing Systems. Red Hook, NY: NeurIPS Foundation, 2024: 17801−17813

[18] Semnani S, Yao V, Zhang Heidi, et al. WikiChat: Stopping the hallucination of large language model chatbots by few-shot grounding on Wikipedia[C]//Proc of the 27th Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2023: 2387−2413

[19] Wang Guangshuo, He Jiajun, Li Hao, et al. RAG-leaks: Difficulty-calibrated membership inference attacks on retrieval-augmented generation[J]. Science China Information Sciences, 2025, 68(6): 160102

[20] Naseh A, Peng Yuefeng, Suri A, et al. Riddle me this! Stealthy membership inference for retrieval-augmented generation[J]. arXiv preprint, arXiv: 2502.00306, 2025

[21] Rossi L, Marek B, Hanke V, et al. Auditing empirical privacy protection of private LLM adaptations[C/OL]//Proc of the 1st Neurips Safe Generative AI Workshop 2024. Red Hook, NY: NeurIPS Foundation, 2024[2025-12-20]. https://openreview.net/forum?id=vFIC8Le5UH

[22] Koga T, Wu Ruihan, Zhang Zhiyuan, et al. Privacy-preserving retrieval-augmented generation with differential privacy[J]. arXiv preprint, arXiv: 2412.04697, 2024

[23] Guu K, Lee K, Tung Z, et al. Retrieval augmented language model pre-training[C]//Proc of the 37th Int Conf on Machine Learning. New York: PMLR, 2020: 3929−3938

[24] Zhang Boyu, Yang Hongyang, Zhou Tianyu, et al. Enhancing financial sentiment analysis via retrieval augmented large language models[C]//Proc of the 4th ACM Int Conf on AI in Finance. New York: ACM, 2023: 349−356

[25] NVIDIA Corporation. NVIDIA ChatRTX: Chat with RTX[R/OL]. 2024[2025-12-20]. https://www. nvidia.com/en-us/ai-on-rtx/chatrtx/

[26] Siriwardhana S, Weerasekera R, Wen E, et al. Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering[J]. Transactions of the Association for Computational Linguistics, 2023, 11: 1−17

[27] Parvez M R, Ahmad W, Chakraborty S, et al. Retrieval augmented code generation and summarization[C]//Proc of the 26th Conf on Empirical Methods in Natural Language Processing (Findings). Stroudsburg, PA: ACL, 2021: 2719−2734

[28] Dwork C, Roth A. The algorithmic foundations of differential privacy[J]. Foundations and Trends® in Theoretical Computer Science, 2014, 9(3/4): 211−407

[29] Abadi M, Chu A, Goodfellow I, et al. Deep learning with differential privacy[C]//Proc of the 23rd ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2016: 308−318

[30] Arachchige P C M, Bertok P, Khalil I, et al. Local differential privacy for deep learning[J]. IEEE Internet of Things J, 2019, 7(7): 5827−5842

[31] Izacard G, Caron M, Hosseini L, et al. Unsupervised dense information retrieval with contrastive learning[J]. arXiv preprint, arXiv: 2112.09118, 2021

[32] Karpukhin V, Ouz B, Min S, et al. Dense passage retrieval for open-domain question answering[C]//Proc of the 25th Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2020: 6769−6781

[33] Cover T M, Hart P E. Nearest neighbor pattern classification[J]. IEEE Transactions on Information Theory, 1967, 13(1): 21−27

[34] Malkov Y A, Yashunin D A. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 42(4): 824−836

[35] Johnson W, Lindenstrauss J. Extensions of lipschitz mappings into a Hilbert space[J]. Contemporary Mathematics, 1984, 26: 189−206

[36] Zoomeye. Ai. ZoomEye[EB/OL]. 2024[2025-12-20]. https://www.zoomeye.org/searchResult?q=aXDpu5Hnmb3lkI3ljZU%3D

[37] Guo M, Dai Zihang, Vrandei D, et al. Wiki-40B: Multilingual language model dataset[C]//Proc of the 12th Language Resources and Evaluation Conf. Marseille: ELRA, 2020: 2440−2452

[38] Liang Zhentao, Mao Jin, Lu Kun, et al. Finding citations for PubMed: A large-scale comparison between five freely available bibliographic data sources[J]. Scientometrics, 2021, 126(12): 9519−9542

[39] Jiang Yi, Yang Yong, Yin Jiali, et al. A survey on security and privacy risks of large language models[J]. Journal of Computer Research and Development, 2025, 62(8): 1979−2018 (in Chinese)
(姜毅, 杨勇, 印佳丽, 等. 大语言模型安全与隐私风险综述[J]. 计算机研究与发展, 2025, 62(8): 1979−2018)

[40] Wang Xiaochen, Zhang Kun, Zhang Peng. A multi-perspective view of large model security and practice[J]. Journal of Computer Research and Development, 2024, 61(5): 1104−1112 (in Chinese)
(王笑尘, 张坤, 张鹏. 多视角看大模型安全及实践[J]. 计算机研究与发展, 2024, 61(5): 1104−1112)



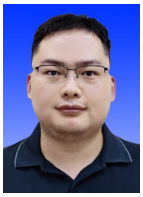
Tan Xue, born in 1997. PhD candidate. His main research interests include large language model security and privacy protection.

谭 学, 1997 年生。博士研究生。主要研究方向为大模型安全、隐私保护。



Zheng Yi, born in 1998. Master. His main research interests include computer vision and larger vision language models.

郑 毅, 1998 年生。硕士。主要研究方向为计算机视觉、视觉语言大模型。



Zhang Yunruo, born in 1996. PhD, postdoc, professor. His main research interests include trusted AI and robustness certification.

张云若, 1996 年生。博士，博士后，研究员。主要研究方向为可信人工智能、模型鲁棒性验证。



Chen Yiran, born in 2002. Master candidate. Her main research interest includes large language model security.

陈怡然, 2002 年生。硕士研究生。主要研究方向为大语言模型安全。



Chen Ping, born in 1985. PhD, professor. His main research interest includes AI security.

陈 平, 1985 年生。博士，研究员。主要研究方向为 AI 安全。



Xue Xiangyang, born in 1968. PhD, professor, PhD supervisor. His main research interests include computer vision, multimedia, and machine learning.

薛向阳, 1968 年生。博士，教授，博士生导师。主要研究方向为计算机视觉、多媒体、机器学习。