ISCAS 中国科学院软件研究所学术年会暨重点实验室科技活动周 2025 第十届
学术论文

RAG-Leaks: Difficulty-Calibrated Membership Inference Attacks on Retrieval-Augmented Generation

Published in Science China Information Sciences, 68, 160102(2025). (CCF-A)

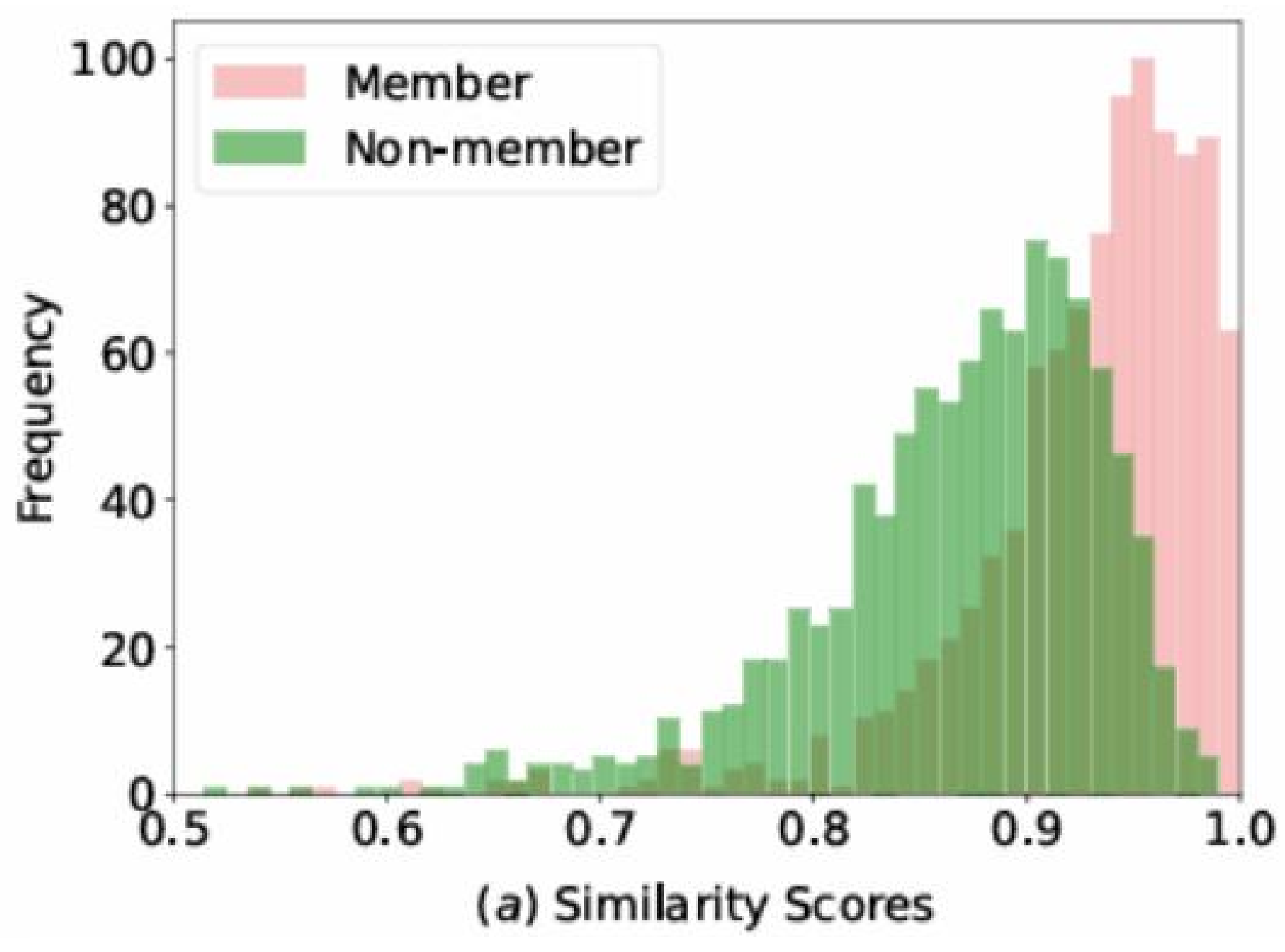
针对检索增强生成的难度校准成员推理攻击

王广硕，何家骏，李昊，张敏，冯登国

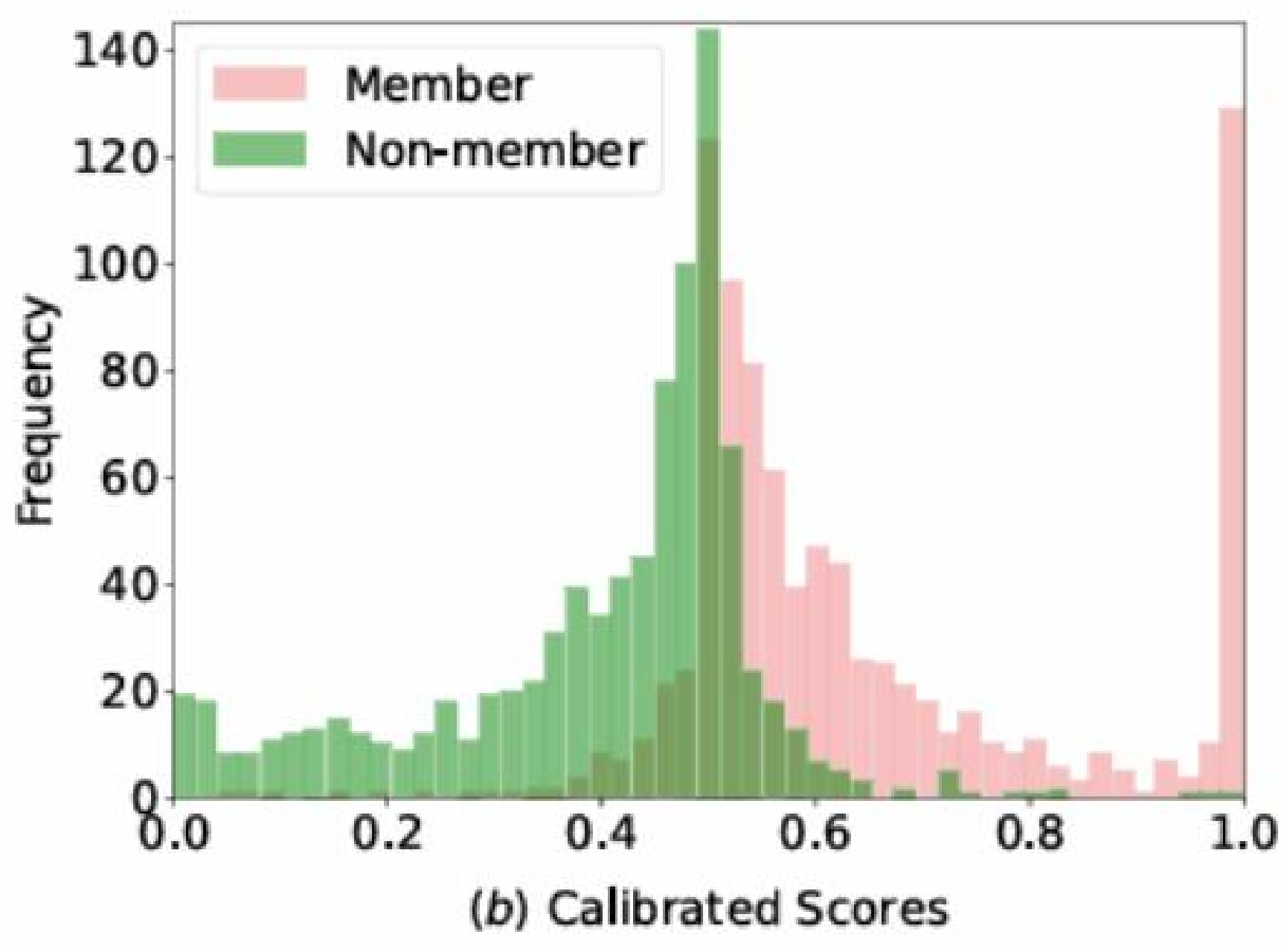
联系人： 李昊 13488718664 lihao@iscas.ac.cn

研究背景

检索增强生成（RAG）是解决大模型（LLM）幻觉问题的一种有效方法，它通过从外部知识库中获取高质量、及时且与上下文相关的信息来增强 LLM 的响应能力。然而，RAG 系统引入的外部知识库包含个人敏感信息，自然引发了对成员推理攻击（MIA）的担忧。作为目前机器学习模型隐私泄露风险检测的主流方式之一，MIA通过观察RAG系统对目标样本的输出来推断“该样本是否属于RAG系统的知识库”。当前，RAG系统中的成员隐私风险仍未得到充分探讨。



(a) Similarity Scores



(b) Calibrated Scores

图1.校准前后的成员信号分布

难度校准

本文首次提出不同的样本在RAG系统中表现出不同程度的回答难度，这是在成员属性和样本多样性的共同作用下引起的。虽然成员属性是一个关键组成部分，但样本多样性的存在会导致低估其隐私风险。例如，图1(a)中成员和非成员在上述两个因素共同作用下展现出较大的重叠，而经过对样本多样性的校准，图1(b)显现了成员和非成员的区分性（即影响因素只剩下成员属性）。

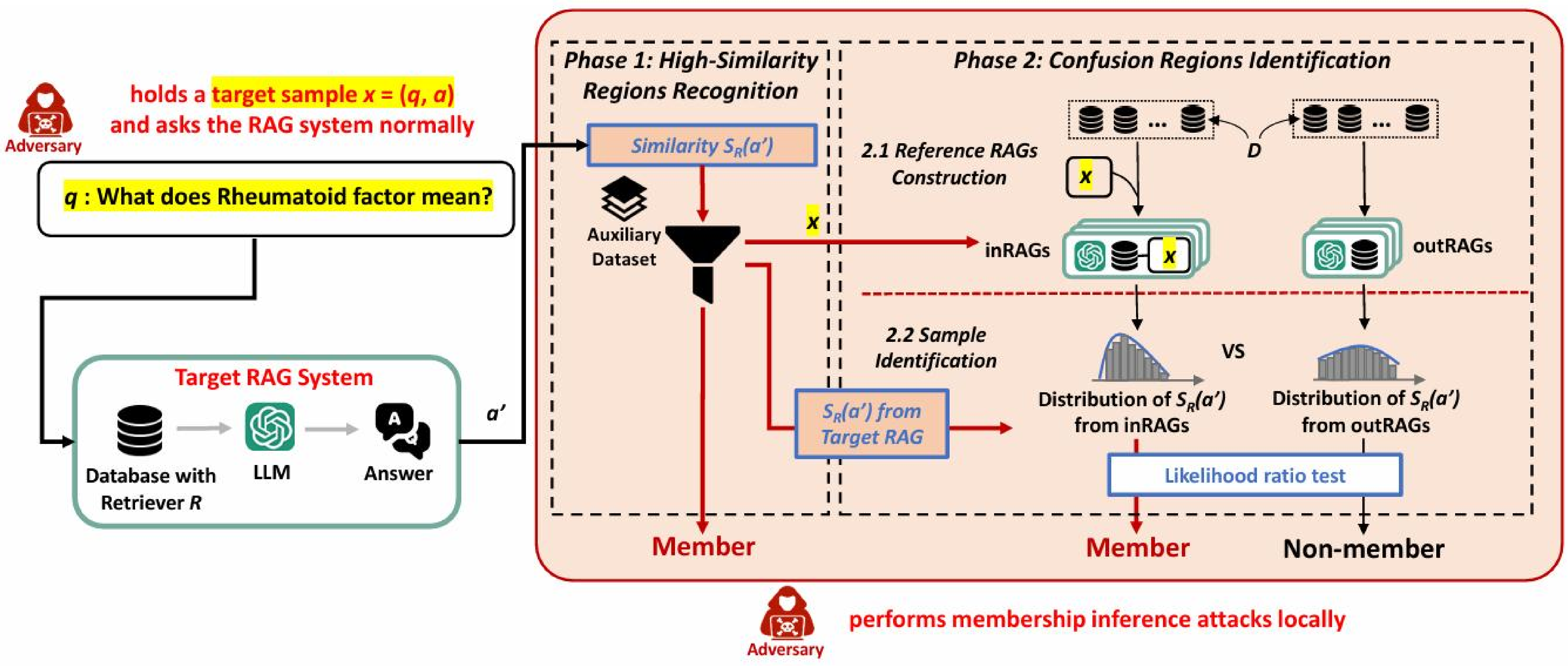


图2.DC-MIA的工作流程

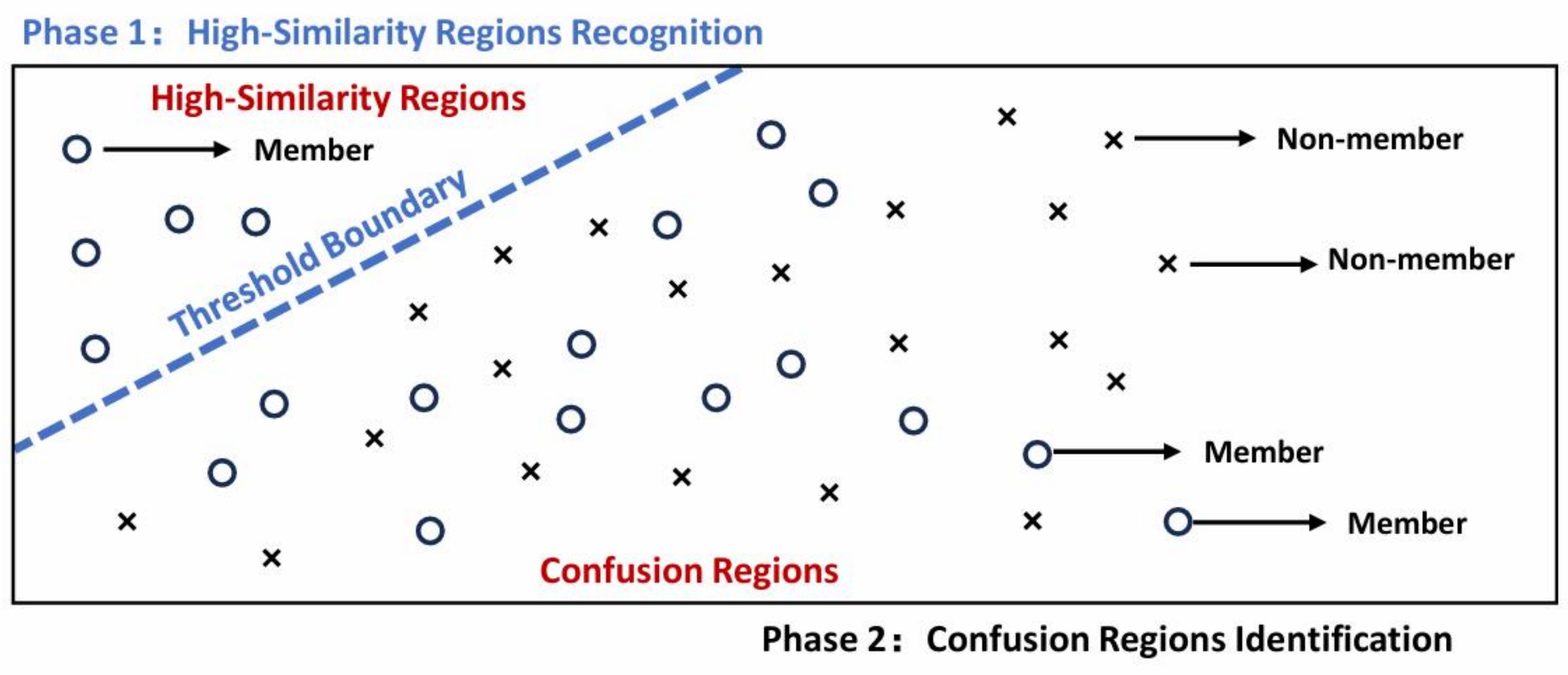


图3.DC-MIA的样本空间分割

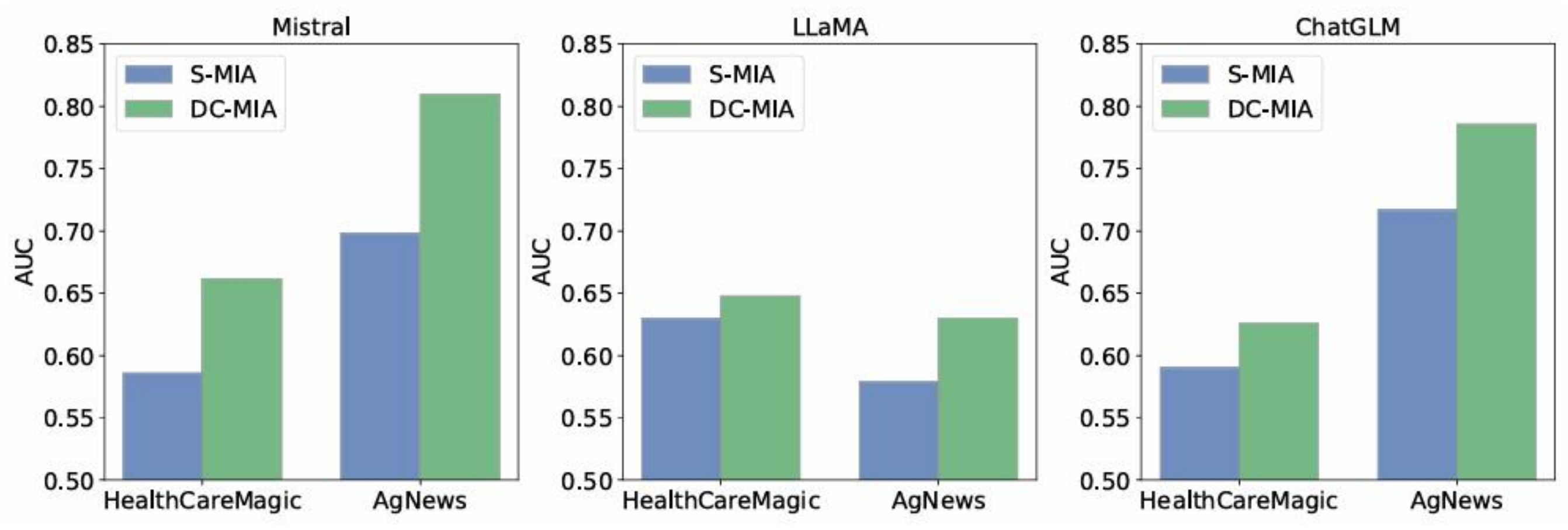


Figure 4: The AUC comparison of DC-MIA and S-MIA attacks on six different RAGs.

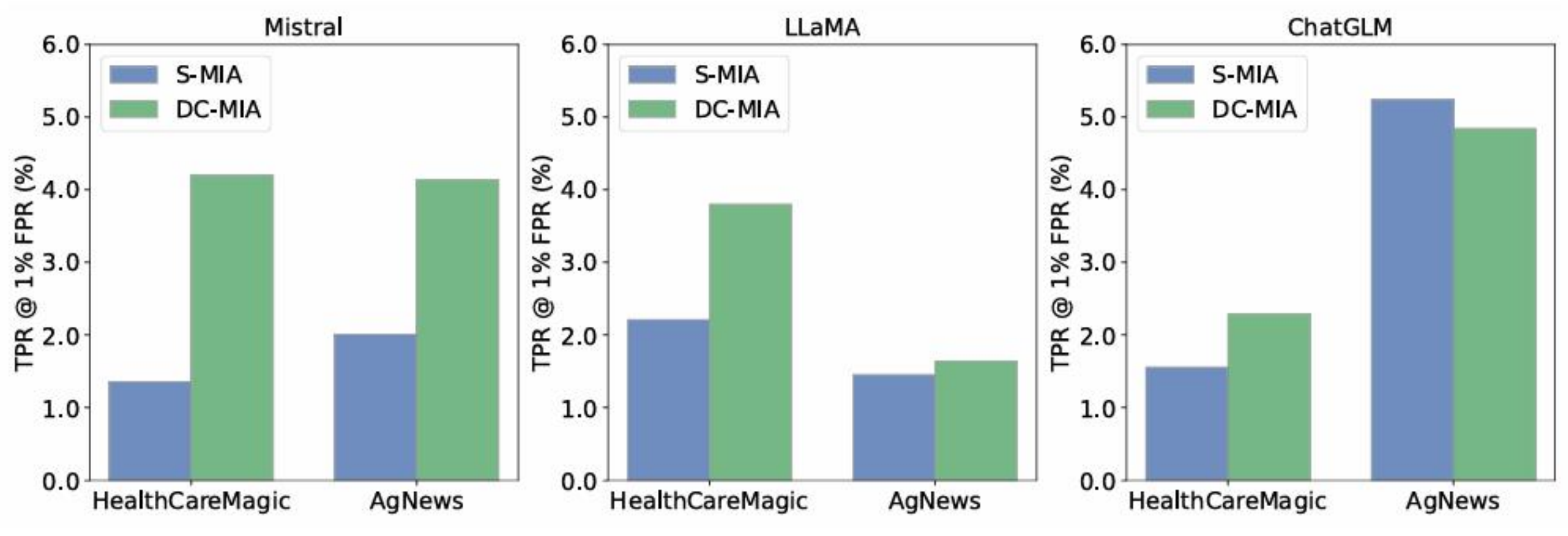


Figure 5: The TPR at 1% FPR comparison of DC-MIA and S-MIA attacks on six different RAGs.

图4.DC-MIA与SOTA方法的性能对比

研究方法

基于难度校正后的信号分布，本文提出了一种针对检索增强生成的成员推理攻击方法DC-MIA。如图3所示，DC-MIA利用辅助数据集计算了一个信号阈值，将样本空间分为两个区域：高相似性区域（主要包含成员）、混淆区域（具有重叠相似性分数的成员和非成员）。对于一个样本，DC-MIA的识别包括两个阶段：

- 高相似性区域识别。如果相似性分数超过阈值，则将其归类为成员，否则划分在混淆区域。
- 混淆区域识别。如图2所示，敌手构建两组具有相同设置的参考RAG系统，inRAG和outRAG，在这两组系统中计算目标样本的相似性分数并分别建立分布。接下来，该样本接受似然比检验，评估其相似性分数更有可能抽样于成员的分布还是非成员的分布。

实验结果

本文在三个基准数据集和三个流行的大型语言模型上进行了广泛的评估。实验结果表明在AUC、TPR@0.1%FPR等指标上，DC-MIA在几乎所有情况下都始终优于SOTA方案。