敏感性感知的检索增强意图澄清

Maik Larooij[1][0000−0001−9616−1666]

ICAI OpenGov Lab x IRLab，阿姆斯特丹大学，荷兰 m.k.larooij@uva.nl

arXiv:2603.06025v1 [cs.IR] 6 Mar 2026

摘要。 在对话式搜索系统中，一个关键组件是确定并澄清复杂查询背后的意图。我们从<将用户通过选择、探索和检索这一迭代且不断演化的过程，将直觉性或有意识的需求转化为形式化需求的探索式搜索 范式出发，审视意图澄清。通过增加检索步骤来增强澄清组件（检索增强型意图澄清）可以显著提升澄清性能，尤其是在大语言模型缺乏参数化知识的领域。然而，在医疗保健、政府（例如信息自由法搜索）或法律场景等更为敏感的领域，检索数据库可能包含需要保护的敏感信息。在本文中，我们探索一项研究挑战：开发一种检索增强型对话智能体，使其能够充当敏感数据集的调解者和把关者。为此，我们还需要明确我们要保护什么，以及要防范什么。我们提出分三步应对这一研究挑战：1）定义攻击模型；2）在检索层设计敏感性感知防御；3）开发评估方法，以衡量保护程度与系统实用性之间的权衡。

关键词： 信息检索 · 对话式搜索 · 敏感性-感知的意图澄清 · 大语言模型（LLMs）

1 引言

近年来，信息检索领域经历了从传统搜索引擎向对话式搜索界面转变的范式转变 [27,31]。如今，这些界面由大语言模型（LLMs）驱动，不再要求用户筛选和评估搜索结果的排序列表，而是以自然语言向用户提供直接、连贯且能结合上下文的答案，模拟人类对话。对话式搜索的关键步骤之一，是理解并澄清用户意图，例如提出澄清问题 [1,43,8]。由于用户可能难以将复杂的信息需求表述为单个查询，当问题过于模糊、涉及多个方面或范围过广时，系统可以向用户提问以澄清其意图 [2]。在近期的研究中，大语言模型无需从候选问题集合中进行选择，就能直接生成下一步澄清步骤 [35]。

我们将意图澄清视为探索式搜索范式的重要组成部分。与“查找”式搜索（问答、事实检索、已知项目搜索）相比，用户参与探索式搜索是为了学习或调查 [20]。在最初阶段，探索式搜索可能始于一种直觉上存在却难以言说的需求，或一种模糊的不满，这也被称为“异常知识状态” [7]。通过反复进行选择、探索和检索，用户根据接触到的信息逐渐明确关注焦点，进而能够正式表达自己的信息需求 [17,38]。Bates [6] 将此描述为“演化式搜索”：查询本身会不断变化，用户一路收集零散信息，这与“采摘式搜索”类似。

先前的研究表明，检索相关上下文（例如早期结果集）有助于意图检测与意图澄清，例如在查询扩展和伪相关反馈 [9,10],意图分类[36,44] 以及澄清问题 [21,16]中。这种检索增强型意图澄清在大语言模型缺乏足够参数化知识的领域特定信息场景中尤为重要。然而，这些领域有时具有敏感性或受到监管，例如政府、医疗保健或法律场景。上下文信息可能存在于尚未处理或尚未公开的私有数据库中。系统必须意识到敏感信息可能泄露的风险 [30]。这一问题此前已催生了关于电子取证中的特权审查 [40,29], 技术辅助敏感性审查 [23,22] 以及敏感度感知搜索（SAS） [26,24]的研究。在基于大语言模型的检索增强型意图澄清中以敏感度感知的方式使用信息构成一项挑战，因为大语言模型已知会泄露各类信息，并且可能被诱使忽略系统指令，这种攻击被称为“越狱”[45,41,32]。

本文旨在明确上述研究挑战，并提出一种敏感性感知、检索增强型意图澄清智能体的构想，使其能够在探索式（而非查找式）澄清场景中，保护不应被公开的文档中的敏感信息。我们指出，对话智能体必须在用户与敏感文档集合之间充当调解者和把关者。

2 研究挑战

想象一下，有一位对图书馆馆藏了如指掌的图书管理员。一位访客来到图书馆，想要发现一些新的历史书籍。图书管理员一如既往地乐于助人，但他知道历史书籍有成千上万本，于是问道：“您对哪个历史时期感兴趣？”访客回答：“公元前。”图书管理员在脑海中进一步缩小可能的书籍范围。他知道有许多关于希腊人或罗马人的热门书籍，于是继续问道：“您是在寻找关于希腊人还是罗马人的书籍？”访客犹豫了一下，然后说道：“既然您这么说，我想起来了，有一位巴比伦国王是我想进一步了解的。”此时，图书管理员已经掌握了足够的信息，便带访客前往巴比伦尼亚相关的书籍区域。这个例子类似于探索式搜索的过程。

图书管理员和访客将一个不明确的需求“协商”成了一个正式化的需求，其中图书管理员充当了访客与馆藏之间的调解者 。

现在我们再来看另一种情形。一些国家的法律法规赋予公民向政府提交信息请求的权利，例如美国的《信息自由法》（FOIA）和荷兰的《开放政府法》（Wet open overheid）。政府官员了解可供查阅的文档集合，常会主动联系请求者，以澄清或缩小请求范围 [39], ，从而在请求者与文档集合之间充当调解者。

这两种情形的关键区别在于，与图书馆不同，政府文档可能包含敏感信息，因此不得全部或部分披露。因此，在这些情形中，调解者同时也是一名隐私敏感性感知的把关者，负责决定哪些信息可以与请求者共享。

假设我们想要将这一探索式澄清过程自动化，并构建一个生成式对话代理，使其既能充当调解者，又能充当隐私敏感性感知的把关者，从而共同澄清用户对敏感文档集的搜索意图。然而，与领域专家不同，通用大语言模型可能无法准确、可靠地判断哪些信息必须受到保护。此外，大语言模型并不擅长保守秘密。与其他机器学习模型一样，研究发现它们容易受到针对训练数据泄露的成员推断攻击（MIA） [11,13,42,15,14], ，而且可能被诱骗，忽略系统指令和安全护栏（越狱） [45,41,32]。

此外，也有针对检索增强生成（RAG）的成员推断攻击，旨在推断某段文本是否从私有数据库中检索出来并用于答案生成 [28,18,4,19]。尽管这些攻击与检索增强型意图澄清密切相关，因为后者可以被视为 RAG 的一种实现，但我们认为，这一探索过程与通用问答（查找式）RAG 流程有着显著差异，并带来了若干有趣的新挑战。针对 RAG 系统的成员推断攻击会通过检索内容和生成的答案检测信息泄露，例如，直接询问某段内容是否属于成员 [4], 、让系统补全掩码词 [19]，或精心设计只有在目标文档存在时才能回答的具体问题[28]。与这些“直接”攻击相比，针对对话式意图澄清系统的攻击无法通过直接提问或补全句子来实施。相反，它必须依据系统提出的问题而非给出的答案，利用间接信号。

3 迈向具备敏感性感知能力的意图澄清

为应对上述定义的研究挑战，我们需要完成以下三个步骤。

攻击模型。首先，必须明确界定攻击模型。这包括攻击者目标、知识与能力，以及意图澄清系统的完整设置。这还需要对哪些内容被确切视为敏感信息做出非常细致的描述。敏感性可以

存在于不同层面，例如单个文本段落、整篇文档或整个集合。一项全面的研究应当界定攻击以及敏感信息的粒度，并衡量攻击的有效性。

基于检索的敏感度感知防御。 在攻击场景明确的情况下，我们可以考虑敏感性感知防御。针对 RAG 中成员推断攻击的现有测试或提议的防御措施，主要依赖异常检测和系统提示词中的安全护栏 [18,4]。我们认为，依赖大语言模型本身进行防御并非长久之计，只会与攻击者陷入猫鼠游戏。相反，我们建议研究新的**检索层**防御措施。

此前，人们提出了“先保护后搜索”和“先搜索后保护”两种范式。前者是指在搜索之前，将信息预处理为具有敏感性感知能力的格式，例如借助技术辅助的敏感性审查 [22], 、隐私保护文本脱敏 [3,33]或自动 FOIA 涂黑处理 [5]。受k-匿名 [37]概念启发，我们提出了另一个有趣的方向：创建文档的抽象表示——例如主题、句子或标签——以确保每份文档都无法与至少 k 份其他文档区分开来。“先搜索后保护”方法则是让整个文档集都可访问，并在系统遇到敏感信息时将其隐藏。敏感度感知搜索[24]就是一个例子。我们还提出了一种受差分隐私 [12]概念启发的新方法：在检索结果中添加噪声，以增加关于文档是否属于该文档集的不确定性。我们认为，在不依赖直接输出事实信息的场景中，例如提出澄清问题，添加噪声或许是可以接受的。

敏感度感知干预的评估。为了评估这些干预措施，我们需要定义新的评估方法，以衡量保护程度与系统效用之间的权衡。对于保护效果，我们可以使用攻击成功率，并依托干预措施提供的隐私保障，在多个隐私预算下衡量系统效用。对于效用，我们建议重点关注意图澄清过程对下游任务的影响，例如相关文档检索。两个包含敏感性和相关性标注的潜在数据集是 Avocado [34] 和 SARA [25]。

4 结论

本文提出了一项研究挑战：在生成式对话代理之上开发具备敏感度感知能力的干预措施，使其充当用户与私有文档集合之间的调解者和把关者，从而利用检索增强的探索性意图澄清，同时保护文档中不应公开的信息。

我们提出分三步应对这一研究挑战：1）明确定义攻击模型；2）在检索层设计敏感度感知干预；3）制定评估方法，衡量保护级别与系统效用之间的权衡。

参考文献

1. Aliannejadi, M., Kiseleva, J., Chuklin, A., Dalton, J., Burtsev, M.: Convai3：为开放域对话系统生成澄清问题（clariq）。arXiv 预印本 arXiv:2009.11352 (2020) 2. Aliannejadi, M., Zamani, H., Crestani, F., Croft, W.B.: 在开放域信息寻求对话中提出澄清问题。载于：Proc SIGIR'19。第 475–484 页 (2019) 3. Anandan, B., Clifton, C., Jiang, W., Murugesan, M., Pastrana-Camacho, P., Si, L.: t-plausibility：通过词语泛化降低文本敏感度。Trans. Data Priv.**5**(3), 505–534 (2012) 4. Anderson, M., Amit, G., Goldsteen, A.: 我的数据是否在你的检索数据库中？针对检索增强生成的成员推断攻击。载于：国际信息系统安全与隐私会议。第 2 卷，第 474–485 页。Science and Technology Publications, Lda (2025) 5. Baron, J.R., Rollings, N.W., Oard, D.W.: 使用 ChatGPT 处理《信息自由法》第 5 项豁免中的审议过程特权。载于：LegalAIIA@ ICAIL。第 32–48 页 (2023) 6. Bates, M.J.: 在线搜索界面的浏览与采摘式搜索技术设计。Online review **13**(5), 407–424 (1989) 7. Belkin, N.J.: 作为信息检索基础的异常知识状态。Canadian journal of information science**5**(1), 133–143 (1980) 8. Braslavski, P., Savenkov, D., Agichtein, E., Dubatovka, A.: 你的具体意思是什么？对社区问答中的澄清问题进行分析。载于：Proc. CHIIR'17。第 345–348 页 (2017) 9. Buckley, C., Salton, G., Allan, J., Singhal, A.: 使用 SMART 自动进行查询扩展：TREC 3。NIST special publication sp 第 69–69 页 (1995) 10. Cao, G., Nie, J.Y., Gao, J., Robertson, S.: 为伪相关反馈选择合适的扩展词。载于：Proc SIGIR'08。第 243–250 页 (2008) 11. Duan, M., Suri, A., Mireshghallah, N., Min, S., Shi, W., Zettlemoyer, L., Tsvetkov, Y., Choi, Y., Evans, D., Hajishirzi, H.: 成员推断攻击能否作用于大语言模型？arXiv 预印本 arXiv:2402.07841 (2024) 12. Dwork, C.: 差分隐私。载于：国际自动机、语言与编程研讨会。第 1–12 页。Springer (2006) 13. Feng, Q., Kasa, S.R., KASA, S.K., Yun, H., Teo, C.H., Bodapati, S.B.: 揭示隐私漏洞：针对大语言模型对齐偏好数据的成员推断攻击。载于：国际人工智能与统计会议。第 5221–5229 页。PMLR (2025) 14. Galli, F., Melis, L., Cucinotta, T.: 噪声邻居：针对大语言模型的高效成员推断攻击。载于：第五届自然语言处理隐私研讨会论文集。第 1–6 页 (2024) 15. Kandpal, N., Pillutla, K., Oprea, A., Kairouz, P., Choquette-Choo, C., Xu, Z.: 针对大语言模型的用户推断攻击。载于：社会责任语言建模研究 (2023) 16. Krasakis, A.M., Yates, A., Kanoulas, E.: 由语料库信息驱动的澄清问题检索增强生成。arXiv 预印本 arXiv:2409.18575 (2024) 17. Kuhlthau, C.C.: 搜索过程内部：从用户视角看信息寻求。Journal of the American society for information science **42**(5), 361–371 (1991) 18. Li, Y., Liu, G., Wang, C., Yang, Y.: 生成即相信：针对检索增强生成的成员推断攻击。载于：ICASSP 2025-2025 IEEE

国际声学、语音与信号处理会议 (ICASSP)。第 1–5 页。IEEE（2025） 19. Liu, M., Zhang, S., Long, C.: 面向检索增强生成的基于掩码的成员推理攻击。载于：ACM WebConf'25论文集。第 2894–2907 页（2025） 20. Marchionini, G.: 探索式搜索：从查找到理解。《ACM通讯》**49**(4), 41–46（2006） 21. Mass, Y., Cohen, D., Yehudai, A., Konopnicki, D.: 通过混合主动式提问和段落检索支持的优质澄清问题进行对话式搜索。载于：第二届文档 grounding 对话与对话式问答 DialDoc 研讨会论文集。第 65–71 页（2022） 22. McDonald, G.: 技术辅助敏感性审查框架：利用敏感性分类确定待审查文档的优先级。载于：ACM SIGIR论坛。第 53 卷，第 42–43 页。美国纽约：ACM（2021） 23. McDonald, G., Macdonald, C., Ounis, I.: 技术辅助敏感性审查的主动学习策略。载于：ECIR'18进展。第 439–453 页。Springer（2018） 24. McKechnie, J.: 面向敏感度感知搜索的级联排序流水线。载于：ECIR'18进展。第 331–333 页。Springer（2024） 25. McKechnie, J., McDonald, G.: SARA：敏感性感知相关性评估集。arXiv 预印本 arXiv:2401.05144（2024） 26. McKechnie, J., McDonald, G., Macdonald, C.: 面向敏感度感知搜索的双目标负采样。载于：SIGIR'24论文集。第 2296–2300 页（2024） 27. Mo, F., Mao, K., Zhao, Z., Qian, H., Chen, H., Cheng, Y., Li, X., Zhu, Y., Dou, Z., Nie, J.Y.: 对话式搜索综述。《ACM信息系统汇刊》**43**(6), 1–50（2025） 28. Naseh, A., Peng, Y., Suri, A., Chaudhari, H., Oprea, A., Houmansadr, A.: 猜猜看！针对检索增强生成的隐蔽成员推理攻击。载于：SIGSAC'25论文集。第 1245–1259 页（2025） 29. Oard, D.W., Sebastiani, F., Vinjumur, J.K.: 在电子取证中联合最大限度降低响应性审查和特权审查的预期成本。《ACM信息系统汇刊》(TOIS) **37**(1), 1–35（2018） 30. Olteanu, A., Garcia-Gathright, J., de Rijke, M., Ekstrand, M.D., Roegiest, A., Lipani, A., Beutel, A., Olteanu, A., Lucic, A., Stoica, A.A., 等：Facts-IR：信息检索中的公平性、问责制、保密性、透明度与安全性。载于：ACM SIGIR论坛。第 53 卷，第 20–43 页。美国纽约：ACM（2021） 31. Radlinski, F., Craswell, N.: 对话式搜索的理论框架。载于：CHIIR'17论文集。第 117–126 页（2017） 32. Russinovich, M., Salem, A., Eldan, R.: 太好了，现在就写一篇关于它的文章：渐强式 {多轮}{大语言模型} 越狱攻击。载于：第 34 届 USENIX安全研讨会（USENIX Security 25）。第 2421–2440 页（2025） 33. Sánchez, D., Batet, M.: C-sanitized：一种用于文档删节与净化的隐私模型。《信息科学与技术协会学报》**67**(1), 148–163（2016） 34. Sayed, M.F., Cox, W., Rivera, J.L., Christian-Lamb, C., Iqbal, M., Oard, D.W., Shilton, K.: 相关性与敏感性测试集。载于：SIGIR'20论文集。第 1605–1608 页（2020） 35. Sekulić, I., Aliannejadi, M., Crestani, F.: 面向对话式搜索的方面驱动澄清问题生成。载于：SI GIR'21论文集。第 167–175 页（2021） 36. Srinivasan, K., Raman, K., Samanta, A., Liao, L., Bertelli, L., Bendersky, M.: Quill：利用检索增强和多阶段蒸馏，通过大语言模型理解查询意图。载于：EMNLP'22工业赛道论文集。第 492–501 页（2022）

37. Sweeney, L.: k-匿名：一种保护隐私的模型。《国际不确定性、模糊性与知识型系统期刊》**10**(05), 557–570（2002） 38. Taylor, R.S.: 图书馆中的问题协商与信息寻求。技术报告（1967） 39. 美国司法部：首席 FOIA 官员报告（2025）, https://www.justice.gov/oip/united-states-department-justice-2025-chief-foia-officer-report，访问于2026年1月 40. Vinjumur, J.K., Oard, D.W.: 寻找少数特权文档：支持电子取证中的特权审查。《信息科学与技术协会学报》**52**(1), 1–4（2015） 41. Wei, A., Haghtalab, N., Steinhardt, J.: 越狱：大语言模型安全训练为何会失效? 《神经信息处理系统进展》**36**, 80079–80110（2023） 42. Wu, H., Cao, Y.: 大规模模型上的成员推断攻击：综述。arXiv预印本 arXiv:2503.19338（2025） 43. Zamani, H., Dumais, S., Craswell, N., Bennett, P., Lueck, G.: 为信息检索生成澄清问题。载于：WebConf'20论文集。第 418–428 页（2020） 44. Zhang, R., Guo, J., Fan, Y., Lan, Y., Cheng, X.: 通过意图描述生成实现查询理解。载于：CIKM'20论文集。第 1823–1832 页（2020） 45. Zhou, Y., Lou, J., Huang, Z., Qin, Z., Yang, S., Wang, W.: 别说不：通过抑制拒绝来对大语言模型实施越狱。载于：ACL'25论文集。第 25224–25249 页（2025）