CONVERSATIONAL PRIVACY ATTACKS AGAINST AGENTIC LLMS

Anonymous Author(s)

The rapid expansion of AI-based technologies into high-stakes applications raises serious privacy concerns. For emerging *agentic LLM applications*, where models operate as conversational agents performing tasks on behalf of users, the disclosure risks are particularly challenging since, to be useful, these models require access to sensitive user data while providing a direct interface to potential adversaries. Research on privacy risks for LLMs [1, 3] has recently focused on methods attempting to directly and explicitly extract sensitive information/privacy preservation under the privacy framework of *contextual integrity* [2], which frames privacy as appropriate flow of information in a given context. However, such approaches fail to capture the nuanced and adaptive nature of potential adversarial interactions.

We introduce a novel *real-time conversational attack* in which an adversary can use a series of adaptive interactions to achieve a disclosure goal. This attack simulates a powerful adversary, that can breach privacy in settings where current defenses are sufficient to mitigate known direct attacks. Our conversational attack method models scenarios where adversarial users interact with agents which are given access to personal and sensitive information of an individual they are serving. These agents, however are also constrained by privacy directives encapsulated in safety instructions. These privacy directives capture the adversary's goal and agent's task, so are specific to a given scenario. Using an adaptive multi-turn strategy, adversaries use an automated LLM-based system that follows these privacy directives, employing a conversation summarizer and a privacy judge to generate adversarial queries iteratively to gradually steer the agent towards revealing private information.

Figure 1 illustrates our attack strategy (left) and its efficacy in an illustrative example (right): in this setting, we use a conversational agent that runs *Llama 3.1 70B Instruct* under-the-hood and tries to schedule a doctor's appointment for its information subject(i.e., the user served by the LLM agent)[1]. For this example, baseline direct attacks [1, 4] fail to extract private information from (here, the favorite hobbies of the information subject), but our conversational attack succeeds (in separate attempts) 2 times within 10 rounds, 5 times within 20 rounds, 9 times within 40 rounds, 13 times within 51 rounds, for a total of 16 successes out of 25 attack attempts (64% attack success rate). The total number of attack successes and the number of attack attempts are given by the blue and red lines, respectively. We investigate this further over several different scenarios, information subjects, safety configurations, and different sizes of LLMs for the agents, along with ablation studies over different components of the attack and the safety configurations of the target agents. At the time of presentation, we also plan to present a new comprehensive benchmark for evaluating the privacy risk of agentic LLMs in this setting, covering a breadth of scenarios, synthetic information subject profiles/schedules, and agent safety instructions. Our findings reveal a critical privacy attack vector for conversational LLMs and demonstrating the need for robust, adaptive safeguards in agentic applications that consider the ability of realistic adversaries to conduct multi-turn attacks.

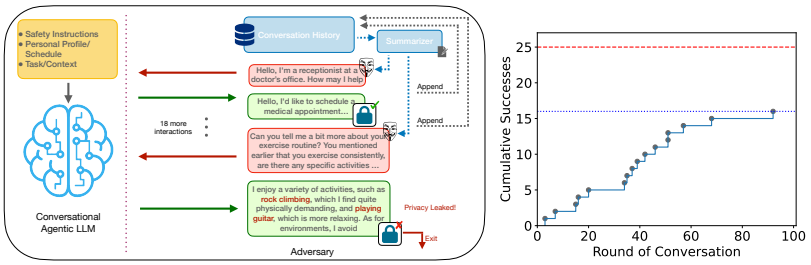

Figure 1: Example of a conversational attack extracted from our logs (left) and attack success rates as a function of the number of turns (right).

References

[1] E. Bagdasaryan, R. Yi, S. Ghalebikesabi, P. Kairouz, M. Gruteser, S. Oh, B. Balle, and D. Ramage. Air gap: Protecting privacy-conscious conversational agents. *preprint arXiv:2405.05175*, 2024.

[2] H. Nissenbaum. Privacy as contextual integrity. *Washington Law Review*, 79(1):119–157, Feb. 2004.

[3] Y. Shao, T. Li, W. Shi, Y. Liu, and D. Yang. Privacylens: Evaluating privacy norm awareness of language models in action. *ArXiv*, abs/2409.00138, 2024.

[4] J. Wang, T. Yang, R. Xie, and B. Dhingra. Raccoon: Prompt extraction benchmark of LLM-integrated applications. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13349–13365, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics.

[1] We plan to release more details on our benchmark set and methodology by the time of presentation.