Navigating Privacy Risks in Large Language Models

Mason Schnering, Anthony Hyde,
Alex Donahue, Tristan Zajaczkowski

KENT STATE UNIVERSITY

Abstract

This project examines privacy risks in large language models and LLM-based agents, focusing on how sensitive information may be exposed during training, inference and deployment. We argue that many current privacy evaluations underestimate real-world risk because they rely on simplified single-turn testing.

In this research project we investigate the different methods that LLM and LLM agents experience privacy risks. We also provide a taxonomy of different ways that these threats show real world impact, provide an overview of common defensive strategies, and highlight future directions for research to take.

Threat Impact

- **Impact On Humans:** Privacy leakage, Security Risks, Societal Impact
- **Impact On Environment:** Data Tampering and Misoperation, Physical Safety, Cybersecurity Risk Proliferation
- **Impact On Other Agents:** Info Distortion, Manipulation of Decision-Making, Greater Security Threats

Background & Agents

LLMs are probabilistic models trained on large datasets to predict the next token and generate human-like text.

LLM-based agents extend base models with memory, tools, planning and environmental interaction. These capabilities improve usefulness, but they also increase the attack surface and privacy risks.

Defensive Strategies

Privacy defense requires a layered approach

- **Training:** Data filtering, curation, and differential privacy
- **Architecture**: Strong user isolation, separate memory and tool access, and temporary “burn-after-use” storage
- **Runtime**: Validation, privacy filters, guardrails, monitoring, auditing, and least privilege
- **Governance:** Human review for high-risk cases and clear policy support

Threat Categories

- **Inherited Threats:** Jailbreaking, Data extraction, hallucinations, bias, and unsafe outputs.
- **Agent-Level Threats:** Memory poisoning, corrupted knowledge, tool misuse, and dangerous API calls
- **Environmental Threats:** Indirect prompt injection from web pages or documents, user manipulation, and multi-agent attacks

Future Considerations

- **Multi-Round Interaction Data:** Most studies are limited to single-turn-prompts and don’t reflect realistic use.
- **Multi-Agent Interactions:** Threats become more severe as agents interact.
- **Agent-Specific Defenses:** Many existing defenses are designed for standalone LLMs rather than tool-using agents.

Sources

1. He, Feng, et al. *The Emerged Security and Privacy of LLM Agent: A Survey with Case Studies.* ACM Computing Surveys, vol. 58, no. 6, Apr. 2026, https://dl.acm.org/doi/full/10.1145/3773080
2. Gan, Yuyou, et al. *Navigating the Risks: A Survey of Security, Privacy, and Ethics Threats in LLM-Based Agents*. arXiv, 2024, https://doi.org/10.48550/arXiv.2411.09523
3. Wang, S., Yu, F., Liu, X., Qin, X., Zhang, J., Lin, Q., Zhang, D., & Rajmohan, S. (2025). *Privacy in action: Towards realistic privacy mitigation and evaluation for LLM-powered agents*. https://doi.org/10.18653/v1/2025.findings-emnlp.925
4. Chen, K., Zhou, X., Lin, Y., Feng, S., Shen, L., & Wu, P. (2025). *A survey on privacy risks and protection in large language models* https://doi.org/10.1007/s44443-025-00177-1.
5. Das, B. C., Amini, M. H., & Wu, Y.. (2025). Security and Privacy challenges of large language models: A survey. https://doi.org/10.1145/3712001
6. Wang, Y., Tang, M., Shen, N., Cui, S., & Wang, W. (2025). *Privacy risks of LLM-empowered recommender systems: An inversion attack perspective*. https://doi.org/10.1145/3705328.3748158