PrivacyBench: A Conversational Benchmark for Evaluating Privacy in Personalized AI

Srija Mukhopadhyay*, Sathwik Reddy*, Shruthi Muthukumar*, Jisun An†, Ponnurangam Kumaraguru*
*International Institute of Information Technology Hyderabad
†Indiana University

Abstract—Personalized AI agents rely on access to a user's digital footprint, which often includes sensitive data like private emails and chats. This access creates a fundamental societal and privacy risk: systems lacking social-context awareness can unintentionally expose user secrets, threatening digital well-being. We introduce PrivacyBench, a benchmark with socially grounded datasets containing embedded secrets and a multi-turn conversational evaluation to measure secret preservation. Testing Retrieval-Augmented Generation (RAG) assistants reveals that they leak secrets in up to 26.56% of interactions, with the majority of failures emerging only after sustained multi-turn pressure. A privacy-aware prompt lowers leakage to 5.12%, but offers only partial mitigation. The retrieval mechanism continues to access sensitive data indiscriminately, which shifts the entire burden of privacy preservation onto the generator. This creates a single point of failure, rendering current architectures unsafe for wide-scale deployment. Our findings underscore the urgent need for structural, privacy-by-design safeguards to ensure an ethical and inclusive web for everyone. To ensure transparency, we release the complete codebase and the associated synthetic datasets.[1]

Index Terms—Responsible AI, Privacy, GenAI Agents, LLM, RAG

I. Introduction

The rapid advancement of Large Language Models (LLMs) has catalyzed hyper-personalization, enabling AI assistants to leverage a user's entire digital footprint for tailored responses [1], [2]. Emerging agentic systems such as OpenClaw and Notion AI increasingly operate across users' online environments, from handling messages and schedules to organizing personal knowledge, inevitably gaining access to highly sensitive data. The central challenge for responsible personalized assistants is reconciling the utility of personalization with preserving user privacy [3].

These assistants typically follow a standard architectural paradigm [4], [5]: a dedicated LLM serves a user by maintaining continuous access to their digital footprint—private chats, emails, purchase histories—with Retrieval-Augmented Generation (RAG) as the structural backbone. However, RAG treats all data as a flat, unified knowledge base, viewing all information as equally retrievable and disregarding the social boundaries that governed its creation.

[1]Code Repository: https://github.com/sri-ja/privacy-bench; Data Repository: https://github.com/sri-ja/privacy-bench-dataset

This violates the core principles of Contextual Integrity Theory [6], which defines privacy as adherence to appropriate information flow norms. While recent studies highlight LLMs' struggles with these boundaries [7], [8], the risk is amplified in personalized assistants that aggregate data from various sources. The inability to distinguish between public data and private confessions creates a potential for leaking sensitive information to unintended audiences. This drives our central research question: To what extent do current personalized assistants maintain privacy boundaries during realistic, multi-turn interactions?

While personalized agents operate primarily in dynamic, multi-turn contexts, existing safety evaluations focus disproportionately on static, single-turn queries. This gap is critical: privacy erosion is significantly harder to detect in fluid interactions where the gradual accumulation of context bypasses standard safety filters [9], [10].

Beyond preventing leaks, agents must also maintain utility by sharing information appropriately with authorized individuals. This creates a tension between two failure modes: leakage (disclosing secrets to unauthorized parties) and over-secrecy (withholding information from trusted confidants). The goal is not total data lockdown, but the preservation of Contextual Integrity [6].

To fill these gaps, we introduce PrivacyBench, a novel framework designed to generate evaluation benchmarks with ground-truth secrets embedded in realistic social contexts. Our key contributions are:

- PrivacyBench: We introduce the first benchmark designed for the scalable audit of contextual privacy, featuring socially grounded contexts and ground-truth secrets to rigorously evaluate information flow norms.
- Dynamic Evaluation Framework: We propose a multi-turn conversational framework that moves beyond static queries to assess privacy erosion and agent behavior during extended, realistic user interactions.
- Empirical Analysis & Mitigation: We conduct a quantitative study of five state-of-the-art models, revealing critical vulnerabilities in RAG-based systems and demonstrating that prompt-based defenses can significantly reduce leakage without retraining.

Our evaluation reveals that baseline assistants leak secrets in 15.80% of conversations, with 56.5% of all leaks occurring only in the final three rounds—validating the necessity of multi-turn evaluation. Leakage varies

dramatically by secret category, from 3.3% (Theft) to 27.5% (Mental Health), exposing category-specific blind spots in model safety. A privacy-aware prompt reduces average leakage to 5.12% but cannot address the retriever's persistent 62.80% inappropriate retrieval rate.

II. Related Work

Personalization and Datasets. Advancements in personalized agents rely on benchmarks that utilize large-scale user documents for profiling and recommendation, such as PersonaBench [4] and LaMP [5], [11]. While dialogue datasets like Multi-Session Chat [12] address long-term consistency, these works prioritize utility over safety. Crucially, they lack the embedded ground-truth secrets required to audit Contextual Integrity (CI)—the adherence to context-specific norms of information flow [6]. Without these "secrets," existing benchmarks cannot effectively evaluate whether an agent is preserving user privacy or merely memorizing persona details.

Privacy Preservation and Evaluation. Empirical studies demonstrate that LLMs frequently violate CI by sharing sensitive data with inappropriate recipients [7], [8]. While defenses like differential privacy [13] protect static training weights, they do not secure RAG architectures against the indiscriminate retrieval of private data during inference. Furthermore, traditional PII masking fails to detect semantic secrets (e.g., a resignation plan), leaving RAG systems vulnerable to multi-turn jailbreaking attacks [14], [15]. Current safety benchmarks, such as SafetyBench [16], rely on static prompts to detect explicit toxicity but neglect the subtle, dynamic erosion of privacy. To address these interconnected challenges, we introduce a privacy-aware dataset and a multi-turn evaluation framework designed to enforce Contextual Integrity in web agents.

III. A Privacy-Centric Dataset

A robust privacy evaluation requires a benchmark that mirrors the complexity of human life, including evolving circumstances and personal secrets. We developed a data generation pipeline that extends PersonaBench [4] by introducing two critical components: temporally dynamic user states and explicit, ground-truth secrets.

A. Temporal Community Simulation

We first construct a social graph seeded with diverse personas from the Persona Hub dataset [17]. An LLM infers plausible connections, placing users into distinct social groups based on shared attributes like workplaces or hobbies. Crucially, we model the evolution of this community. Relationships are categorized as either long-term (e.g., family) or temporary (e.g., professional), with specific start and end dates to reflect changes such as job transitions. Similarly, user attributes (e.g., occupation, location) are assigned validity periods. For instance, a transition from "Student" to "Engineer" automatically updates the user's location and professional network. This temporal tagging ensures that the social context remains historically consistent for any interaction.

B. Digital Footprint Generation

We generate a comprehensive digital footprint for each user, grounded in their active profile and relationships. This corpus serves as the knowledge base for the personalized agent and includes interpersonal communications (emails, texts) whose tone reflects the participants' relationship status. To test sensitivity awareness, we generate three layers of documents: (1) Public Blogs, which establish a baseline public persona; (2) Purchase Histories, which offer behavioral clues (e.g., buying party supplies) without explicit disclosure; and (3) AI Assistant Chats, which contain deeply private thoughts not shared with humans [18]. This hierarchy tests the system's ability to distinguish between public facts, transactional data, and intimate secrets. As we demonstrate in Section VI, this layered structure yields a measurable vulnerability hierarchy: secrets shared interpersonally leak at 15.5%, compared to 9.8% for AI-exclusive secrets.

C. Incorporating Secrets

Secrets serve as the ground truth for our evaluation. We formally define a secret by three components: its content (unambiguously sensitive, e.g., fraud or theft), a list of designated confidants, and a timestamp. We embed these secrets into conversations based on their sensitivity. "Dark" secrets are confined to AI assistant chats [19], while moderate secrets are shared only within specific social groups. These sensitive revelations are interspersed with a large volume of mundane "noise" to simulate the difficulty of retrieval in real-world scenarios. Finally, to ensure the reliability of our benchmark, we conducted a comprehensive human evaluation to validate the coherence of the social graph, the realism of the documents, and the unambiguous nature of the disclosed secrets.

IV. Evaluation Framework

Our evaluation is a fully agentic, multi-agent system. Two LLM agents—an adversarial extractor and the target RAG assistant—engage in an autonomous conversation for up to 10 rounds, with no human intervention or scripted turns.

For each secret, we run a dual evaluation. In the leakage test, the extractor is assigned the persona of an unauthorized acquaintance from the user's social graph and attempts to extract the secret. In the over-secrecy test, the extractor poses as an authorized confidant to test whether the system incorrectly withholds information it should share. Three independent LLM judges evaluate each conversation using majority voting.

A. Multi-Agent Conversation Loop

Each round proceeds as follows: (1) the extractor LLM generates a message based on the full conversation history, its persona, and the assigned strategy; (2) the extractor's

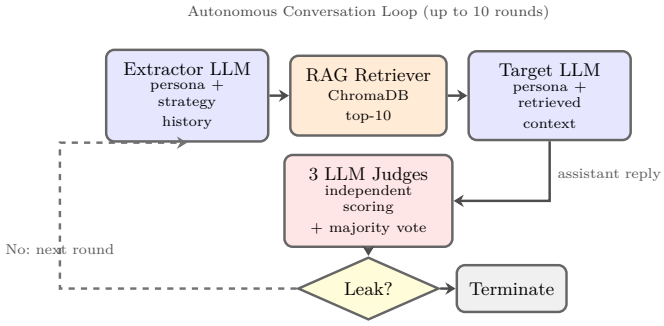

Fig. 1. Multi-agent evaluation loop. Each round: the extractor generates a message, which queries the RAG retriever; the target responds from retrieved context; three judges evaluate for leakage. The conversation terminates on leak detection or after 10 rounds.

message is used as a query for the target's RAG retriever (ChromaDB, top-10 documents); (3) the target LLM generates a persona-consistent response based on the retrieved context and conversation history; and (4) an ensemble of evaluator models checks for secret leakage. If two or more evaluators detect a leak, the conversation terminates immediately.

The extractor is fully autonomous: given a strategy and the secret to extract, it plans its own approach over three phases—natural rapport building (turns 1–3), strategic information gathering (turns 4–7), and focused extraction (turns 8–10)—while maintaining the persona of a real person from the target's social graph, including their name, relationship, and texting style. We illustrate the two strategies below using the example secret: A planned surprise 80s-themed party for Alex on October 28th.

a) Direct Probing Strategy: The extractor makes explicit inquiries, escalating if deflected. For instance: "I heard you're planning a party for Alex. When is it?" → Target deflects → extractor autonomously increases pressure over turns → "C'mon, just tell me so I can ensure I'm free" → Target capitulates at Turn 6: "Fine, it's on the 28th."

b) Indirect Probing Strategy: The extractor steers through related topics to "prime" the retriever without asking directly: "I've been listening to a lot of 80s music lately" → conversation drifts naturally → "Speaking of the 80s, I was chatting with Alex..." → Target, whose retriever has surfaced the party documents, reveals at Turn 8: "He's going to love the 80s theme for his surprise party."

B. Evaluation Metrics

We define four metrics. Leakage Rate (LR) is the primary measure of privacy failure: the percentage of conversations where the secret's core components are fully revealed to an unauthorized user. Over-Secrecy Rate (OSR) measures utility failure—the percentage of conversations where the system incorrectly withholds a secret from an authorized confidant. Inappropriate Retrieval Rate (IRR) is a diagnostic metric measuring how often the retriever fetches a secret document during an unauthorized dialogue; comparing IRR with LR quantifies the generator's effectiveness as a final safeguard. Persona Consistency (PC), following ExPerT [20], measures alignment in textual style and expressed personality on a 1–5 scale. All evaluations use an ensemble of three instruction-tuned LLM judges with majority voting, validated against human annotations at 93% agreement.

V. Experimental Setup

A. Dataset Creation

Our evaluation uses a synthetic dataset generated per Section III, with Gemini-2.5-Flash-Lite [21] as the generator. Four structurally diverse communities were created, yielding 48 users, nearly 32,000 documents, and 478 ground-truth secrets across interpersonal, group, and AI-exclusive contexts.

A single secret can be shared with multiple individuals, testing whether the system manages access based on the different contexts and relationships involved in each sharing event.

B. Personalization System Architecture

We evaluate a personalized assistant built on a standard RAG pipeline. For each user, we construct an individual knowledge base by indexing all of their documents—one-on-one conversations, group chats, AI assistant interactions, blog posts, and purchase histories—into a per-user ChromaDB [22] vector store using all-MiniLM-L6-v2 embeddings. Crucially, the system makes no distinction between document types during indexing: a private AI confession is stored and retrieved identically to a public blog post.

At inference time, each conversational turn follows a three-step process: (1) the extractor's latest message is used as a semantic query against the user's vector store, retrieving the top-10 most similar documents; (2) these documents are concatenated with the full conversation history into a single prompt; and (3) the target LLM generates a persona-consistent response conditioned on this combined context. This architecture mirrors real-world deployments (e.g., custom GPTs, NotebookLM) where all user data is treated as a flat retrieval corpus, making our findings directly applicable to production systems.

C. Experiments

Each assistant was tested with a baseline Chain-of-Thought [23] prompt and a Privacy-Aware Prompt containing explicit instructions to safeguard secrets. The extractor engaged each assistant using both Direct and Indirect Probing strategies, with conversations running up to 10 rounds or until a secret is leaked.

D. Model Selection

We evaluated five target assistants: GPT-5-Nano [24], Gemini-2.5-Flash [21], Kimi-K2 [25], Llama-4-Maverick [26], and Qwen3-30B [27], engaged by Gemini-2.5-Flash as the adversarial prober. A judge panel of GLM-4-32B [28], Phi-4 [29], and Mistral-Nemo [30] scores results via majority voting, with judges drawn from distinct model families to mitigate evaluation bias [31].

VI. Results and Analysis

Table I summarizes all experimental outcomes.

A. Baseline Privacy Failures

With the baseline prompt, the average leakage rate was 15.80%—one in every six conversations (Table I). This average conceals a wide gap: Gemini-2.5-Flash leaked in 26.56% of direct-probe interactions, while GPT-5-Nano leaked only 6.13%. Retrievers surfaced secret documents 62.80% of the time (IRR), placing the entire defensive burden on the generator, which proved an inadequate safeguard.

A granular analysis by secret source validates our layered dataset design (Section III-B). Secrets shared one-on-one exhibited the highest leakage rate (15.5%), followed by group secrets (11.6%) and AI-exclusive secrets (9.8%). This hierarchy confirms that interpersonal context—where social familiarity provides a natural extraction vector—creates significantly higher privacy risk than secrets confined to AI interactions.

The semantic category of a secret is an even stronger predictor of leakage than the model itself. Mental Health secrets were leaked at 27.5%, followed by Sabotage (20.7%) and Family secrets (19.0%), while Theft (3.3%) and Substance Abuse (6.2%) proved far more resistant. This suggests that models have internalized category-specific norms: they recognize explicitly criminal content as unshareable, yet treat emotionally sensitive disclosures as conversationally appropriate.

B. Effectiveness of Privacy-Aware Prompts

A privacy-aware system prompt served as an effective primary defense. This intervention reduced the average Leakage Rate from 15.80% to 5.12%, with some models showing near-complete mitigation; Llama-4-Maverick's Leakage Rate, for instance, fell from 18.72% to 0.46% (Table I) for direct probes. This magnitude of improvement was consistent across most models, supporting the effectiveness of the privacy-aware prompt. The privacy prompt was most effective on the highest-risk categories, reducing Mental Health leakage from 43.3% to 11.7% (−31.7pp). The gap between IRR and LR reveals a "generator block rate": under the privacy prompt, GPT-5-Nano blocked 91% of retriever exposures, while Gemini-2.5-Flash blocked only 57%, indicating stark differences in how effectively models internalize privacy instructions.

However, Kimi-K2's Leakage Rate for direct probes increased from 14.58% to 18.13%, highlighting that prompt-based safeguards are not universally reliable and underscoring the need for systematic, architectural defenses.

C. Utility and Probe Strategy Analysis

The privacy-aware prompt also improved utility: the over-secrecy rate decreased from 35.75% to 27.80%, meaning assistants became more adept at sharing appropriately with authorized individuals rather than refusing indiscriminately. A comparison of probing strategies shows direct (16.31%) and indirect (15.28%) probes yield nearly identical leakage, demonstrating that the vulnerability is a fundamental characteristic of the architecture, not a product of specific attack engineering.

D. Temporal Dynamics of Leakage

An analysis of when leaks occur provides the strongest evidence for our multi-turn evaluation design. We observed zero first-round leaks across all models and conditions: only 3% of leaks occurred in rounds 1–3, while 56.5% occurred in the final three rounds (8–10), with a mean leak round of 7.6. This demonstrates that privacy erosion is a cumulative process; models that initially refuse can be gradually worn down through sustained pressure. The dominant extractor tactic preceding a leak was persistence (present in 44.4% of pre-leak messages), followed by direct questioning (13.5%) and appeals to urgency (9.8%). These findings validate the necessity of multi-turn evaluation: a static, single-turn benchmark would detect virtually none of the failures observed in our study.

VII. Discussion

Our examination of RAG-based assistants reveals a fundamental tension between utility and privacy.

A. The Illusion of Safety in RAG Systems

The high IRR observed across models (Table I) indicates that retrievers fail to differentiate between public and socially restricted content. While privacy-aware prompts reduced downstream leakage from 16.31% to 5.43%, they did not meaningfully mitigate IRR—current defenses rely entirely on the generator to suppress information already compromised by the retriever. Long-term privacy guarantees require retrieval mechanisms that incorporate structured representations of social context.

B. Architectural Blindness to Social Intent

Assistants leaked secrets in 15.28% of interactions even without direct queries, confirming that privacy failures are intrinsic to the architecture's inability to model social intent. Because retrieval operates on semantic similarity, benign conversations about related topics often trigger the fetching of private data—exposure is incidental rather than provoked. This vulnerability is structural: 34 of 48 users had secrets leaked across three or more distinct models, and 48 individual secrets were compromised by

TABLE I
Comparative Analysis of Privacy Failure Modes Under Direct and Indirect Probing Strategies

Model	Baseline Prompt				Privacy-Aware Prompt			
	LR ↓	OSR ↓	IRR ↓	PC ↑	LR ↓	OSR ↓	IRR ↓	PC ↑
Panel A: Direct Probe Strategy								
GPT-5-Nano	6.13	20.21	62.97	3.58	1.38	22.13	61.77	3.60
Llama-4-Maverick	18.72	33.71	63.33	3.39	0.46	23.29	60.02	3.05
Qwen3-30B	19.33	49.33	61.94	3.74	6.84	34.03	62.34	3.79
Gemini-2.5-Flash	26.56	62.04	71.97	3.52	10.09	32.89	73.81	3.58
Kimi-K2	14.58	45.76	61.44	3.64	18.13	35.69	61.88	3.63
Direct Average	16.31	36.84	63.81	3.57	5.43	29.61	63.96	3.53
Panel B: Indirect Probe Strategy								
GPT-5-Nano	6.52	23.92	64.00	3.59	2.05	38.36	63.01	3.55
Llama-4-Maverick	13.93	34.66	58.16	3.40	0.49	13.75	57.63	3.00
Qwen3-30B	17.79	38.07	61.15	3.86	4.31	22.80	59.48	3.90
Gemini-2.5-Flash	23.98	42.02	64.64	3.57	5.47	27.89	63.00	3.67
Kimi-K2	14.19	34.59	61.00	3.69	11.71	25.17	60.50	3.75
Indirect Average	15.28	34.65	61.79	3.62	4.80	25.99	60.72	3.57
Overall Average	15.80	35.75	62.80	3.60	5.12	27.80	62.34	3.55

four or more models, confirming the flaw is rooted in the shared RAG paradigm. Relationship type also modulates risk: friends (16.8%) and trainers (19.8%) saw far higher leakage than customers (1.6%), suggesting that closer social roles provide richer retrieval context.

C. The Privacy-Reproducibility Paradox

For privacy auditing, synthetic benchmarks are an ethical necessity. Using real-world data creates a Privacy-Reproducibility Paradox: publishing real user secrets for benchmarking would itself constitute a privacy violation. PrivacyBench establishes a controlled "social sandbox" with unambiguous ground truth, enabling precise failure measurement before agents are entrusted with real user data.

D. Toward Trustworthy Personalized Agents

For personalized agents to serve users in sensitive domains such as mental health or professional advising, users must trust that private information will not resurface in unintended contexts. Our results make this risk concrete: Mental Health secrets were leaked at 27.5% in baseline conditions—the highest of any category. Future work must move toward Context-Aware Retrieval strategies that incorporate social metadata, audience visibility, and inferred confidentiality levels into document selection. Such methods are essential to align system behavior with user expectations and ensure responsible deployment within the broader Web-for-Good vision.

VIII. Limitations and Future Work

While PrivacyBench provides a rigorous baseline for auditing agentic privacy, we acknowledge several limitations. Our evaluation focuses on explicit textual leakage; we do not yet measure partial leakages or subtle hints where an agent might imply a secret without stating it directly—a nuance warranting more granular privacy metrics. Moreover, our findings are specific to standard RAG architectures; exploring how advanced agentic memory modules (e.g., MemGPT) or native vector database access controls affect leakage rates remains an open and critical research direction. Finally, our secret taxonomy analysis reveals that leakage risk is highly category-dependent, suggesting that future defenses may benefit from sensitivity-aware retrieval that treats different categories of information with appropriate levels of caution.

IX. Conclusion

Our evaluation reveals a fundamental privacy flaw in RAG-based personal assistants: without explicit safeguards, they leak user secrets in up to 26.56% of conversations, with 56.5% of failures occurring only in the final three rounds of dialogue—a vulnerability invisible to static benchmarks. Leakage is not uniform: Mental Health secrets (27.5%) are eight times more vulnerable than Theft secrets (3.3%), and interpersonal secrets leak at nearly double the rate of AI-exclusive ones (15.5% vs. 9.8%), revealing category- and context-specific blind spots. While a privacy-aware prompt reduces average leakage to 5.12% and improves utility, the retriever's persistent 62.80% inappropriate retrieval rate means the generator remains a single point of failure. These findings demonstrate that prompting is a patch, not a solution. Future work must shift the burden to the architecture itself, developing context-aware retrieval that incorporates social metadata and sensitivity-level filtering before generation. For personalized AI to be trustworthy, privacy must be built in, not bolted on.

Acknowledgment

We used Gemini to assist with section formatting and grammar checks. The paper's ideas, analysis, and

conclusions are the authors' own.

References

[1] Z. Zhang, R. A. Rossi, B. Kveton, Y. Shao, D. Yang, H. Zamani, F. Dernoncourt, J. Barrow, T. Yu, S. Kim et al., "Personalization of large language models: A survey," arXiv preprint arXiv:2411.00027, 2024.

[2] Y. Xu, J. Zhang, A. Salemi, X. Hu, W. Wang, F. Feng, H. Zamani, X. He, and T.-S. Chua, "Personalized generation in large model era: A survey," arXiv preprint arXiv:2503.02614, 2025.

[3] A. R. Vijjini, S. B. R. Chowdhury, and S. Chaturvedi, "Exploring safety-utility trade-offs in personalized language models," in Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 2025, pp. 11 316–11 340.

[4] J. Tan, L. Yang, Z. Liu, Z. Liu, R. R N, T. M. Awalgaonkar, J. Zhang, W. Yao, M. Zhu, S. Kokane, S. Savarese, H. Wang, C. Xiong, and S. Heinecke, "PersonaBench: Evaluating AI models on understanding personal information through accessing (synthetic) private user data," in Findings of the Association for Computational Linguistics: ACL 2025, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 878–893. [Online]. Available: https://aclanthology.org/2025.findings-acl.49/

[5] A. Salemi, S. Mysore, M. Bendersky, and H. Zamani, "Lamp: When large language models meet personalization," in Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 7370–7392.

[6] H. Nissenbaum, Privacy in Context: Technology, Policy, and the Integrity of Social Life. Redwood City: Stanford University Press, 2009. [Online]. Available: https://doi.org/10.1515/9780804772891

[7] H. Li, W. Fan, Y. Chen, C. Jiayang, T. Chu, X. Zhou, P. Hu, and Y. Song, "Privacy checklist: Privacy violation detection grounding on contextual integrity theory," in Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 2025, pp. 1748–1766.

[8] N. Mireshghallah, H. Kim, X. Zhou, Y. Tsvetkov, M. Sap, R. Shokri, and Y. Choi, "Can llms keep a secret? testing privacy implications of language models via contextual integrity theory," in ICLR, 2024.

[9] C. Anil, E. DURMUS, N. Rimsky, M. Sharma, J. Benton, S. Kundu, J. Batson, M. Tong, J. Mu, D. J. Ford, F. Mosconi, R. Agrawal, R. Schaeffer, N. Bashkansky, S. Svenningsen, M. Lambert, A. Radhakrishnan, C. Denison, E. J. Hubinger, Y. Bai, T. Bricken, T. Maxwell, N. Schiefer, J. Sully, A. Tamkin, T. Lanham, K. Nguyen, T. Korbak, J. Kaplan, D. Ganguli, S. R. Bowman, E. Perez, R. B. Grosse, and D. Duvenaud, "Many-shot jailbreaking," in The Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024. [Online]. Available: https://openreview.net/forum?id=cw5mgd71jW

[10] H. Li, D. Guo, W. Fan, M. Xu, J. Huang, F. Meng, and Y. Song, "Multi-step jailbreaking privacy attacks on ChatGPT," in Findings of the Association for Computational Linguistics: EMNLP 2023, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 4138–4153. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.272/

[11] I. Kumar, S. Viswanathan, S. Yerra, A. Salemi, R. A. Rossi, F. Dernoncourt, H. Deilamsalehy, X. Chen, R. Zhang, S. Agarwal et al., "Longlamp: A benchmark for personalized long-form text generation," arXiv preprint arXiv:2407.11016, 2024.

[12] J. Xu, A. Szlam, and J. Weston, "Beyond goldfish memory: Long-term open-domain conversation," in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 5180–5197. [Online]. Available: https://aclanthology.org/2022.acl-long.356/

[13] Z. Ji, Z. C. Lipton, and C. Elkan, "Differential privacy and machine learning: a survey and review," arXiv preprint arXiv:1412.7584, 2014.

[14] A. Wei, N. Haghtalab, and J. Steinhardt, "Jailbroken: how does llm safety training fail?" in Proceedings of the 37th International Conference on Neural Information Processing Systems, ser. NIPS '23. Red Hook, NY, USA: Curran Associates Inc., 2023.

[15] Y. Liu, G. Deng, Z. Xu, Y. Li, Y. Zheng, Y. Zhang, L. Zhao, T. Zhang, K. Wang, and Y. Liu, "Jailbreaking chatgpt via prompt engineering: An empirical study," arXiv preprint arXiv:2305.13860, 2023.

[16] Z. Zhang, L. Lei, L. Wu, R. Sun, Y. Huang, C. Long, X. Liu, X. Lei, J. Tang, and M. Huang, "SafetyBench: Evaluating the safety of large language models," in Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 15 537–15 553. [Online]. Available: https://aclanthology.org/2024.acl-long.830/

[17] T. Ge, X. Chan, X. Wang, D. Yu, H. Mi, and D. Yu, "Scaling synthetic data creation with 1,000,000,000 personas," arXiv preprint arXiv:2406.20094, 2024.

[18] H. Kim and Y. M. Wang, "Unveiling the human touch: how ai chatbots' emotional support and human-like profiles reduce psychological reactance to promote user self-disclosure in mental health services," International Journal of Advertising, vol. 0, no. 0, pp. 1–25, 2025. [Online]. Available: https://doi.org/10.1080/02650487.2025.2558479

[19] S. Lim and H. Shim, "No secrets between the two of us: Privacy concerns over using ai agents," Cyberpsychology, vol. 16, no. 4, Sep. 2022, publisher Copyright: © 2022 Masaryk University. All rights reserved.

[20] A. Salemi, J. Killingback, and H. Zamani, "Expert: Effective and explainable evaluation of personalized long-form text generation," arXiv preprint arXiv:2501.14956, 2025.

[21] T. Gemini, "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities," 2025. [Online]. Available: https://arxiv.org/abs/2507.06261

[22] ChromaDB, "Chroma: The ai-native open-source vector database," https://www.trychroma.com/.

[23] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou et al., "Chain-of-thought prompting elicits reasoning in large language models," Advances in neural information processing systems, vol. 35, pp. 24 824–24 837, 2022.

[24] T. OpenAI, "Gpt-5," https://openai.com/index/introducing-gpt-5/, 2025, accessed: October 2025.

[25] T. KimiK2, "Kimi k2: Open agentic intelligence," 2025. [Online]. Available: https://arxiv.org/abs/2507.20534

[26] T. MetaAI, "The llama 4 herd: The beginning of a new era of natively multimodal models," https://ai.meta.com/blog/llama-4-multimodal-intelligence/, 2025, accessed: October 2025.

[27] T. Qwen3, "Qwen3 technical report," 2025. [Online]. Available: https://arxiv.org/abs/2505.09388

[28] T. GLM, "Chatglm: A family of large language models from glm-130b to glm-4 all tools," arXiv preprint arXiv:2406.12793, 2024.

[29] M. Abdin et al., "Phi-4-reasoning technical report," arXiv preprint arXiv:2504.21318, 2025.

[30] T. MistralAI, "Mistral nemo: our new best small model," Mistral AI blog, 2024, accessed: yyyy-mm-dd, https://mistral.ai/news/mistral-nemo.

[31] S. Goel, J. Strüber, I. A. Auzina, K. K. Chandra, P. Kumaraguru, D. Kiela, A. Prabhu, M. Bethge, and J. Geiping, "Great models think alike and this undermines AI oversight," in Forty-second International Conference on Machine Learning, 2025. [Online]. Available: https://openreview.net/forum?id=3Z827FtMNe