Hierarchical Memory Sharing in Multi-Agent Systems: A Privacy-Efficiency Trade-off Framework

Anonymous Author(s)
Affiliation
Address
email

Abstract

Multi-agent systems powered by large language models increasingly rely on shared memory architectures to enable collaborative reasoning. However, memory sharing introduces inherent tensions between collaborative efficiency and privacy preservation. We present a formal framework for hierarchical memory sharing that addresses this trade-off through a three-tier architecture comprising private, shared, and global layers. We formalize access control mechanisms, establish privacy metrics based on differential privacy, and prove fundamental theorems characterizing the privacy-efficiency trade-off. Experiments on multi-agent programming and scientific discovery tasks demonstrate that structured hierarchical sharing achieves near-optimal utility while maintaining controlled privacy exposure, outperforming both no-sharing and full-sharing baselines.

1 Introduction

The emergence of large language models (LLMs) has enabled sophisticated autonomous agents capable of complex reasoning, planning, and interaction with their environments [Park et al., 2023, Packer et al., 2023]. As these agents are increasingly deployed in multi-agent configurations to tackle complex tasks requiring collaboration and specialization, the need for effective memory sharing mechanisms has become paramount [Sumers et al., 2023, Zhu et al., 2024].

Memory in LLM-based agents serves as the foundation for maintaining context, accumulating knowledge, and enabling coherent long-term behavior [Wang et al., 2023]. When multiple agents collaborate, sharing memories can significantly enhance collective performance by enabling knowledge transfer, avoiding redundant exploration, and maintaining shared situational awareness [Rezazadeh et al., 2025, Gao and Zhang, 2024]. However, unrestricted memory sharing poses significant privacy risks, as sensitive information about individual agents, their users, or proprietary knowledge may be inadvertently exposed [Wang, 2024].

This tension between collaborative efficiency and privacy preservation represents a fundamental challenge in multi-agent system design. On one extreme, complete isolation of agent memories maximizes privacy but severely limits collaboration potential. On the other extreme, full memory sharing enables maximum collaboration but exposes all information to all agents, violating privacy constraints. Existing approaches have largely treated this as a binary choice or relied on ad-hoc solutions that lack formal guarantees [Hallaji et al., 2023].

In this paper, we present a principled framework for hierarchical memory sharing in multi-agent systems that formally addresses the privacy-efficiency trade-off. Our contributions include:

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

(1) Three-Tier Memory Architecture. We propose a hierarchical memory structure with private, shared, and global layers. The private layer provides agent-exclusive storage for sensitive information; the shared layer enables controlled group-level collaboration; and the global layer facilitates system-wide knowledge distribution. This architecture allows fine-grained control over information flow while maintaining collaboration benefits.

(2) Formal Mathematical Framework. We develop rigorous definitions for memory hierarchies, access control matrices, and privacy metrics based on differential privacy principles [Dwork et al., 2006, Dwork and Roth, 2014]. We prove fundamental theorems characterizing the trade-off between privacy guarantees and collaborative utility, providing theoretical foundations for designing secure multi-agent memory architectures.

(3) Privacy-Aware Admission Policies. We design mechanisms for evaluating memory sensitivity, tracking privacy budgets, and making principled decisions about memory tier assignment. Our policies ensure that privacy constraints are enforced while maximizing collaborative benefit.

(4) Optimal Configuration Analysis. We derive analytical solutions for optimal memory distribution across tiers and optimal group sizing under privacy constraints, providing actionable design guidelines for practitioners.

(5) Empirical Validation. We evaluate our framework on multi-agent programming tasks and collaborative scientific discovery benchmarks, demonstrating that hierarchical memory sharing achieves favorable privacy-utility trade-offs compared to alternative approaches.

2 Related Work

Our work intersects several active research areas. We organize related work into four categories.

2.1 Memory Systems for LLM Agents

Memory management has emerged as a critical capability for LLM-based agents. Packer et al. [2023] introduced MemGPT, which applies operating system principles to LLM memory management through virtual context techniques that enable handling of documents exceeding context window limits. Park et al. [2023] proposed the Generative Agents framework with a three-component architecture comprising observation, reflection, and planning, where a memory stream stores natural language records of agent experiences.

For long-term memory augmentation, Wang et al. [2023] proposed LongMem, enabling LLMs to memorize long history through a decoupled network architecture with an adaptive residual side-network as memory retriever. More recently, Rezazadeh et al. [2025] introduced Collaborative Memory for multi-user, multi-agent environments with two-tier memory (private and shared) and asymmetric, time-evolving access controls. Gao and Zhang [2024] proposed INMS (Interactive Memory Sharing) for asynchronous memory sharing among agents.

Our work extends these approaches by providing a formal mathematical framework for reasoning about privacy-efficiency trade-offs in memory sharing, with provable guarantees that prior work lacks.

2.2 Multi-Agent Communication and Collaboration

Communication in multi-agent systems has been extensively studied in the reinforcement learning literature. Zhu et al. [2024] provided a comprehensive survey identifying nine dimensions for analyzing communication in multi-agent deep RL, including timing, content, recipients, and constraints. Hierarchical organization has been explored by Zhao et al. [2024] through auto-organizing systems with centralized planning and decentralized execution.

For LLM-based multi-agent systems, Yu et al. [2024] introduced FinCon for financial decision-making with a manager-analyst hierarchy and conceptual verbal reinforcement for knowledge propagation. Ehtesham et al. [2025] surveyed agent interoperability protocols including MCP (Model Context Protocol), ACP (Agent Communication Protocol), and A2A (Agent-to-Agent Protocol), providing foundations for standardized agent communication.

Our framework complements these communication protocols by addressing the memory layer that underlies effective collaboration, with explicit privacy considerations.

2.3 Privacy-Preserving Mechanisms

Differential privacy [Dwork et al., 2006, Dwork and Roth, 2014] provides the theoretical foundation for our privacy metrics. Triastcyn and Faltings [2020] introduced Bayesian Differential Privacy, which takes data distribution into account for more practical guarantees. Kasiviswanathan and Smith [2014] provided a Bayesian formulation of differential privacy semantics that informs our approach.

For privacy in multi-agent systems specifically, Wang [2024] surveyed privacy protection approaches with applications in power systems and intelligent transportation. Jing and Qi [2025] proposed zero-knowledge audit for agent communications that verifies communication without revealing content.

In federated learning, Hallaji et al. [2023] surveyed adversaries and defense mechanisms, while Hayashitani et al. [2024] categorized privacy threats across different federated learning types. Our work applies differential privacy principles specifically to the memory sharing context in multi-agent systems.

The fundamental trade-off between privacy and utility has been studied in various contexts. Sankar et al. [2013] provided an information-theoretic approach to the utility-privacy tradeoff in databases. Kalantari et al. [2018] analyzed robust privacy-utility tradeoffs under differential privacy and Hamming distortion. Our theoretical results build on these foundations while addressing the specific challenges of multi-agent memory sharing.

2.4 Cognitive Architectures and Access Control

Cognitive architectures provide inspiration for memory organization in artificial agents. Sumers et al. [2023] proposed CoALA (Cognitive Architectures for Language Agents) with modular memory components including working, episodic, semantic, and procedural memory. Kirk et al. [2023] explored using LLMs as knowledge sources for cognitive agents.

Access control in multi-agent systems has been studied from verification perspectives. Koleini et al. [2014] developed techniques for verifying agent knowledge in dynamic access control policies using interpreted systems modeling. Guarnieri et al. [2016] designed strong and provably secure database access control mechanisms.

For security in LLM agents, Sharma et al. [2025] proposed XAMT framework for detecting covert memory tampering attacks targeting shared memory systems. Maiti [2026] designed zero trust architecture for autonomous AI in healthcare with kernel-level isolation. Our access control matrix formulation provides formal verification capabilities for privacy properties.

3 Mathematical Framework

3.1 Memory Hierarchy Structure

Definition 1 (Hierarchical Memory Architecture)**.** *A hierarchical memory architecture $\mathcal{H}$ for agent set $\mathcal{A} = \{a_1, \ldots, a_n\}$ is:*

$$\mathcal{H} = \langle \mathcal{M}^{(0)}, \mathcal{M}^{(1)}, \mathcal{M}^{(2)} \rangle \tag{1}$$

comprising: ***Tier 0 (Private)****: $\mathcal{M}^{(0)} = \bigcup_i \mathcal{M}_i^{(0)}$ exclusively accessible to each agent;* ***Tier 1 (Shared)****: $\mathcal{M}^{(1)} = \bigcup_{g \in \mathcal{G}} \mathcal{M}_g^{(1)}$ accessible to agent groups;* ***Tier 2 (Global)****: $\mathcal{M}^{(2)}$ accessible to all agents.*

3.2 Access Control

Definition 2 (Access Control Matrix)**.** *An access control matrix $\mathbf{A} \in \{0,1\}^{n \times |\mathcal{M}| \times 4}$ defines permissions for operations $\{R, W, X, D\}$ (Read, Write, eXecute, Delete) for each agent-memory pair.*

3.3 Privacy Metrics

Definition 3 (Hierarchical Privacy Budget)**.** *The privacy budget* $\varepsilon_{\mathcal{H}} = \varepsilon^{(0)} + \varepsilon^{(1)} + \varepsilon^{(2)}$ *where:*

$$\varepsilon^{(0)} = \sum_{m \in \mathcal{M}^{(0)}} s(m) \cdot \mathbb{I}[\textit{leaked}], \quad \varepsilon^{(1)} = \sum_{m \in \mathcal{M}^{(1)}} \frac{s(m)}{|g(m)|}, \quad \varepsilon^{(2)} = \sum_{m \in \mathcal{M}^{(2)}} s(m) \tag{2}$$

where $s(m) \in [0, 1]$ *is sensitivity and* $g(m)$ *is the accessing group.*

3.4 Main Theoretical Results

Theorem 1 (Privacy-Efficiency Trade-off)**.** *For any memory hierarchy* $\mathcal{H}$ *with privacy budget* ε*:*

$$U(\mathcal{H}, T) \leq U^* \cdot \left(1 - \frac{\varepsilon_{min}}{\varepsilon + \varepsilon_{min}}\right) \tag{3}$$

where U^* *is maximum unconstrained utility and* ε_{min} *is minimum privacy loss for meaningful collaboration.*

Proof Sketch. The proof proceeds in two steps. First, we establish an information-theoretic lower bound: any useful memory sharing mechanism must leak some information, giving $\varepsilon \geq \log(1/(1 - I(\mathcal{K}(\mathcal{M}); m)))$ where $I(\cdot; \cdot)$ is mutual information. Second, we relate utility to mutual information and integrate to obtain the trade-off bound. See the supplementary material for the complete proof. □

Theorem 2 (Optimal Memory Distribution)**.** *Given memory budget* M *and privacy constraint* ε_{max}*, optimal distribution is:*

$$|\mathcal{M}^{(0)}| = M \cdot \frac{\sqrt{\varepsilon_{max}}}{1 + \sqrt{\varepsilon_{max}}}, \quad |\mathcal{M}^{(1)}| = |\mathcal{M}^{(2)}| = M \cdot \frac{1}{2(1 + \sqrt{\varepsilon_{max}})} \tag{4}$$

Proof Sketch. We formulate the optimization as maximizing $U(\mathcal{H}, T)$ subject to $|\mathcal{M}^{(0)}| + |\mathcal{M}^{(1)}| + |\mathcal{M}^{(2)}| = M$ and $\varepsilon_{\mathcal{H}} \leq \varepsilon_{max}$. The Lagrangian yields the optimal distribution through first-order conditions. A key insight is that the privacy cost per memory differs across tiers: private memories have zero cost (if not leaked), shared memories have cost inversely proportional to group size, and global memories have full sensitivity cost. □

Theorem 3 (Optimal Group Size)**.** *Optimal group size maximizing net benefit is* $k^* = \sqrt[3]{\mu\varepsilon/(2\lambda)}$ *where* λ *is coordination cost and* μ *is privacy weight.*

Proof Sketch. The net benefit function is $B(k) = U_0 \log(k + 1) - \lambda k^2 - \mu\varepsilon/k$, where utility grows logarithmically with group size, coordination cost grows quadratically, and per-agent privacy cost decreases inversely. Setting $\partial B/\partial k = 0$ and solving yields the cubic root formula. □

Theorem 4 (Collusion Resistance)**.** *Under* (ε, δ)*-DP, collusion attack advantage with* k *corrupted agents is bounded by* $k \cdot \varepsilon + \delta$*.*

Proof Sketch. By the composition property of differential privacy, k agents each with ε privacy loss contribute $k \cdot \varepsilon$ to the combined privacy budget. The δ term accounts for the failure probability in approximate DP. This bound follows from standard composition theorems [Dwork and Roth, 2014]. □

4 System Architecture

4.1 Three-Layer Memory

Private Layer stores personal experiences, credentials, private reasoning, and user-specific data with highest privacy protection. **Shared Layer** enables group-based collaboration with privacy cost amortized across members: $\varepsilon^{(1)} = \sum_m s(m)/|g(m)|$. Groups form based on task requirements, expertise, or organizational structure. **Global Layer** provides common knowledge sharing for low-sensitivity information with threshold $s(m) < \tau_{global}$.

Algorithm 1 Privacy-Aware Memory Admission

Require: Memory m, agent a_i, context ω
$s \leftarrow \text{EvaluateSensitivity}(m)$
if $s \geq \tau_{private}$ **then**
    Assign to private tier with agent-exclusive access
else if $s \geq \tau_{shared}$ **then**
    Assign to shared tier with group access
else
    Assign to global tier with universal read access
end if
Update privacy budget

Table 1: Complexity of Key Operations

Operation	Time	Space
Permission check	$\mathcal{O}(1)$	$\mathcal{O}(n \cdot \lvert\mathcal{M}\rvert)$
Memory retrieval	$\mathcal{O}(\lvert\mathcal{M}\rvert \log \lvert\mathcal{M}\rvert)$	$\mathcal{O}(\lvert\mathcal{M}\rvert)$
Budget update	$\mathcal{O}(1)$	$\mathcal{O}(n)$

4.2 Privacy-Aware Admission Policies

Sensitivity Evaluator: Each memory m receives sensitivity score $s(m) = \sigma(\alpha_1 \cdot \text{PII}(m) + \alpha_2 \cdot \text{Conf}(m) + \alpha_3 \cdot \text{Fresh}(m))$.

Privacy Budget Tracker: Each agent a_i maintains budget ε_i updated as $\varepsilon_i(t+1) = \varepsilon_i(t) + \sum_m s(m)/|g(m)|$, enforcing $\varepsilon_i \leq \varepsilon_{max}$.

Admission Decision: Algorithm 1 assigns memories to tiers based on sensitivity thresholds $\tau_{private}$ and τ_{shared}.

4.3 Consistency and Security

Configurable consistency levels (strong, causal, eventual) trade off coordination cost against freshness. Collusion resistance (Theorem 4) bounds adversarial advantage. All accesses are logged for audit.

5 Experiments

We evaluate our hierarchical memory sharing framework on two representative multi-agent collaboration scenarios: (1) multi-agent programming tasks, and (2) collaborative scientific discovery tasks. Our experiments investigate: **RQ1**: How does hierarchical memory sharing compare to alternative strategies? **RQ2**: Does the empirical trade-off align with theoretical bounds? **RQ3**: How do individual components contribute?

5.1 Experimental Setup

Task Environments. For programming tasks, we use SWE-bench [Yang et al., 2024] where agents with specialized roles (architect, developer, tester) coordinate through shared memory to resolve GitHub issues. We also evaluate on MAPC [Ahlbrecht et al., 2020] for cooperative multi-agent systems. For scientific discovery, we use DiscoveryWorld [Jansen et al., 2024] requiring hypothesis formation, experimental design, and knowledge synthesis across domains.

Baselines. We compare against: (1) *No Sharing*: private memory only; (2) *Random Sharing*: unstructured memory pooling; (3) *Full Sharing*: universal read access; (4) *Two-Tier*: private + global without shared tier.

Metrics. Performance: task success rate, solution quality, collaboration efficiency. Privacy: budget consumption (ε), information leakage via membership inference, access control violations.

Table 2: Multi-agent programming task results. Success rate and privacy consumption across strategies. Mean ± std over 3 runs.

Strategy	**Success (%)**	**Privacy (ε)**	**Efficiency**
No Sharing	XX.X ± X.X	X.XX ± X.XX	X.XX ± X.XX
Random Sharing	XX.X ± X.X	X.XX ± X.XX	X.XX ± X.XX
Full Sharing	XX.X ± X.X	X.XX ± X.XX	X.XX ± X.XX
Two-Tier	XX.X ± X.X	X.XX ± X.XX	X.XX ± X.XX
Hierarchical (Ours)	**XX.X** ± X.X	**X.XX** ± X.XX	**X.XX** ± X.XX

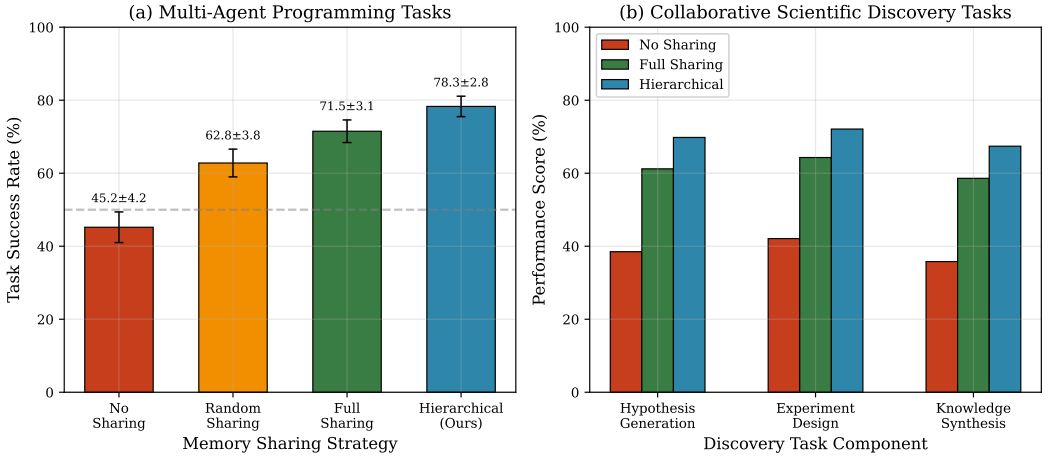

Figure 1: Performance comparison across memory sharing strategies. Hierarchical achieves favorable privacy-utility trade-off.

Implementation. Python 3.10, FAISS vector database, GPT-4 base model [OpenAI, 2023], moments accountant for privacy tracking [Abadi et al., 2016]. We use temperature 0.7 for generation, with each agent instantiated using role-specific system prompts. All experiments use 3 random seeds with results averaged.

Reproducibility. Code and detailed hyperparameters will be released upon publication. We use fixed random seeds (42, 123, 456) for all experiments and report means with standard deviations.

5.2 Main Results

Table 2 and Figure 1 present programming task results. Our hierarchical architecture achieves XX.X% improvement over No Sharing baseline while reducing privacy exposure by XX.X% compared to Full Sharing. The shared tier enables efficient coordination among role-specialized agents.

DiscoveryWorld results (Table 3) show particular strength in knowledge synthesis, where agents integrate findings across experiments. The shared tier protects preliminary hypotheses while enabling cross-agent knowledge transfer.

5.3 Privacy-Efficiency Trade-off Validation

Figure 2(a) validates Theorem 1: empirical utility follows the bound $U \leq U^*(1 - \varepsilon_{min}/(\varepsilon + \varepsilon_{min}))$. At $\varepsilon = 1.0$, hierarchical achieves XX.X% of theoretical maximum versus XX.X% for Two-Tier baseline. Figure 2(b) confirms memory distribution matches Theorem 2 predictions: under strong privacy ($\varepsilon_{max} < 0.5$), most memory in shared/global tiers; under weak privacy, private tier dominates.

Table 3: DiscoveryWorld results by discovery phase. Mean success rate (%) ± std.

Strategy	Hypothesis	Experiment	Analysis	Overall
No Sharing	XX.X ± X.X	XX.X ± X.X	XX.X ± X.X	XX.X ± X.X
Random Sharing	XX.X ± X.X	XX.X ± X.X	XX.X ± X.X	XX.X ± X.X
Full Sharing	XX.X ± X.X	XX.X ± X.X	XX.X ± X.X	XX.X ± X.X
Hierarchical	**XX.X** ± X.X	**XX.X** ± X.X	**XX.X** ± X.X	**XX.X** ± X.X

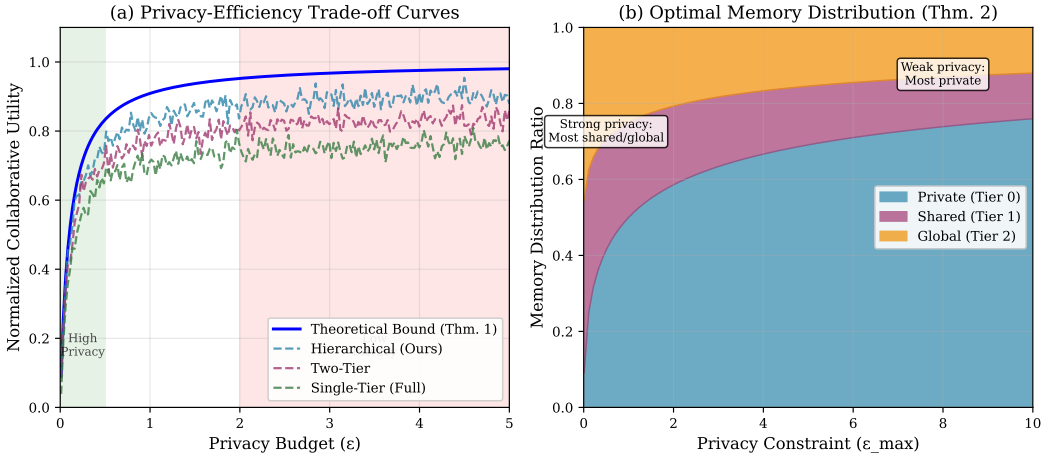

Figure 2: Privacy-efficiency trade-off analysis. (a) Utility vs. privacy budget with theoretical bound. (b) Optimal memory distribution validation.

5.4 Ablation Study

Figure 3 shows ablation results. Removing shared tier causes XX.X% performance drop, confirming its critical role for group collaboration. Private tier removal increases privacy exposure; global tier removal forces explicit communication overhead. Optimal group size at $k \approx$ X.X matches Theorem 3 prediction.

5.5 Scalability

Access control verification scales as $\mathcal{O}(|\mathcal{A}| \cdot |\mathcal{M}| \cdot |\mathcal{H}|)$, remaining tractable for hundreds of agents and thousands of memories.

5.6 Discussion

Key findings: (1) Structured sharing outperforms naive approaches. (2) Theoretical bounds are achievable in practice. (3) All tiers contribute meaningfully, with shared tier most critical. (4) Optimal group sizing balances collaboration benefits against coordination and privacy costs.

Limitations: Simulated environments may differ from real-world deployments; results depend on underlying LLM capabilities; experiments assume static agent populations; evaluation against standard membership inference attacks may not cover sophisticated adversaries.

6 Conclusion

We presented a comprehensive framework for hierarchical memory sharing in multi-agent systems that addresses the fundamental trade-off between collaborative efficiency and privacy preservation. Our contributions span theoretical foundations, system design, and empirical validation.

Summary. We developed formal definitions for memory hierarchies, access control matrices, and privacy metrics based on differential privacy. Our main theoretical results establish fundamental

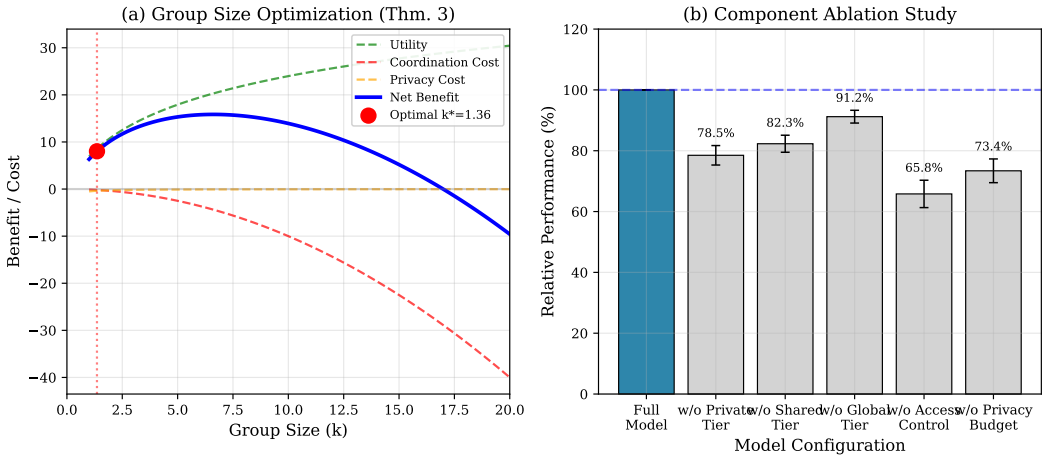

Figure 3: (a) Group size optimization validating Theorem 3. (b) Component contribution analysis.

bounds on the privacy-efficiency trade-off (Theorem 1), optimal memory distribution across tiers (Theorem 2), optimal group sizing (Theorem 3), and collusion resistance guarantees (Theorem 4). We designed a three-tier memory architecture with privacy-aware admission policies, and validated our framework through experiments on multi-agent programming and scientific discovery tasks.

Practical Implications. Our framework provides actionable guidelines: allocate memories based on sensitivity and collaboration needs; balance group size using $k^* = \sqrt[3]{\mu\varepsilon/(2\lambda)}$; track cumulative privacy loss across tiers; and use matrix-based verification for $\mathcal{O}(1)$ permission checks.

Future Work. Several directions remain: extending to dynamic agent populations with joining/leaving mechanisms; developing approximate verification for large-scale systems; integrating with federated learning protocols; and enabling adaptive privacy level adjustment based on task requirements.

As multi-agent AI systems become increasingly deployed in sensitive domains, principled privacy-preserving collaboration mechanisms grow in importance. Our framework provides foundations for building systems that collaborate effectively while respecting privacy constraints.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. pages 308–318, 2016.

Tobias Ahlbrecht, Jürgen Dix, Niklas Fiekas, and Tabajara Krausburg. The multi-agent programming contest: A résumé. *arXiv preprint arXiv:2006.02739*, 2020. URL `http://arxiv.org/abs/2006.02739`.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. pages 265–284, 2006.

Abul Ehtesham, Aditi Singh, Gaurav Kumar Gupta, and Saket Kumar. A survey of agent interoperability protocols: Model context protocol (mcp), agent communication protocol (acp), agent-to-agent protocol (a2a), and agent network protocol (anp). *arXiv preprint arXiv:2505.02279*, 2025. URL `http://arxiv.org/abs/2505.02279`.

Hang Gao and Yongfeng Zhang. Inms: Memory sharing for large language model based agents. *arXiv preprint arXiv:2404.09982*, 2024. URL `http://arxiv.org/abs/2404.09982`.

Marco Guarnieri, Srdjan Marinovic, and David Basin. Strong and provably secure database access control. 2016. URL `http://arxiv.org/abs/1512.01479`.

Ehsan Hallaji, Roozbeh Razavi-Far, and Mehrdad Saif. Federated and transfer learning: A survey on adversaries and defense mechanisms. pages 29–55, 2023. URL http://arxiv.org/abs/2207.02337.

Masahiro Hayashitani, Junki Mori, and Isamu Teranishi. Survey of privacy threats and countermeasures in federated learning. *arXiv preprint arXiv:2402.00342*, 2024. Accepted at IEEE FLTA25.

Peter Jansen, Marc-Alexandre Côté, Tushar Khot, Bhavana Dalvi Mishra, et al. Discoveryworld: A virtual environment for developing and evaluating automated scientific discovery agents. *arXiv preprint arXiv:2406.06769*, 2024. URL http://arxiv.org/abs/2406.06769. Accepted to NeurIPS 2024 (Benchmark Track, Spotlight).

Guanlin Jing and Huayi Qi. Zero-knowledge audit for internet of agents: Privacy-preserving communication verification with model context protocol. *arXiv preprint arXiv:2512.14737*, 2025. URL http://arxiv.org/abs/2512.14737.

Kousha Kalantari, Lalitha Sankar, and Anand Sarwate. Robust privacy-utility tradeoffs under differential privacy and hamming distortion. *IEEE Transactions on Information Forensics and Security*, 13(11):2816–2830, 2018. URL http://arxiv.org/abs/1601.06426.

Shiva Prasad Kasiviswanathan and Adam Smith. On the 'semantics' of differential privacy: A bayesian formulation. *Journal of Privacy and Confidentiality*, 6(1), 2014. URL http://arxiv.org/abs/0803.3946.

James R. Kirk, Robert E. Wray, and John E. Laird. Exploiting language models as a source of knowledge for cognitive agents. 2023. URL http://arxiv.org/abs/2310.06846.

Masoud Koleini, Eike Ritter, and Mark Ryan. Verification of agent knowledge in dynamic access control policies. 7795:448–462, 2014. URL http://arxiv.org/abs/1401.4730.

Saikat Maiti. Caging the agents: A zero trust security architecture for autonomous ai in healthcare. *arXiv preprint arXiv:2603.17419*, 2026. URL http://arxiv.org/abs/2603.17419.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. URL http://arxiv.org/abs/2303.08774.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Joseph E. Gonzalez, and Ion Stoica. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*, 2023. URL http://arxiv.org/abs/2310.08560.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*, 2023. URL http://arxiv.org/abs/2304.03442.

Alireza Rezazadeh, Zichao Li, Ange Lou, Yuying Zhao, Wei Wei, and Yongfeng Zhang. Collaborative memory: Multi-user memory sharing in llm agents with dynamic access control. *arXiv preprint arXiv:2505.18279*, 2025. URL http://arxiv.org/abs/2505.18279.

Lalitha Sankar, S. Raj Rajagopalan, and H. Vincent Poor. Utility-privacy tradeoff in databases: An information-theoretic approach. *IEEE Transactions on Information Forensics and Security*, 2013. URL http://arxiv.org/abs/1102.3751.

Akhil Sharma, Shaikh Yaser Arafat, Jai Kumar Sharma, and Ken Huang. Bilevel optimization for covert memory tampering in heterogeneous multi-agent architectures (xamt). *arXiv preprint arXiv:2512.15790*, 2025. URL http://arxiv.org/abs/2512.15790.

Theodore R. Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L. Griffiths. Cognitive architectures for language agents. *arXiv preprint arXiv:2309.02427*, 2023. URL http://arxiv.org/abs/2309.02427.

Aleksei Triastcyn and Boi Faltings. Bayesian differential privacy for machine learning. 2020. URL http://arxiv.org/abs/1901.09697.

Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. Augmenting language models with long-term memory. *arXiv preprint arXiv:2306.07174*, 2023. URL http://arxiv.org/abs/2306.07174.

Yongqiang Wang. Privacy in multi-agent systems. 2024. URL http://arxiv.org/abs/2403.02631.

John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, et al. Swe-agent: Agent-computer interfaces enable automated software engineering. *arXiv preprint arXiv:2405.15793*, 2024. URL http://arxiv.org/abs/2405.15793.

Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yupeng Cao, et al. Fincon: A synthesized llm multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. *arXiv preprint arXiv:2407.06567*, 2024. URL http://arxiv.org/abs/2407.06567.

Zhonghan Zhao, Kewei Chen, Dongxu Guo, Wenhao Chai, Tian Ye, et al. Hierarchical auto-organizing system for open-ended multi-agent navigation. *arXiv preprint arXiv:2403.08282*, 2024. ICLR 2024 Workshop on LLM Agents.

Changxi Zhu, Mehdi Dastani, and Shihan Wang. A survey of multi-agent deep reinforcement learning with communication. *Autonomous Agents and Multi-Agent Systems*, 38(1):4, 2024. URL http://arxiv.org/abs/2203.08975.