FROM LOCAL LEAKAGE TO GLOBAL PRIVACY RISK IN SEQUENTIAL MULTI-AGENT SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sequential multi-agent large language model (LLM) systems are increasingly deployed in sensitive domains such as healthcare, finance, and enterprise decision-making, where multiple specialized agents collaboratively process a single user request. Although individual agents may satisfy local privacy constraints, sensitive information can still be inferred through sequential composition and intermediate representations. In this work, we study *compositional privacy leakage* in sequential LLM agent pipelines. We formalize leakage using mutual information and derive a compositional upper bound relating agent-local leakage to final-output privacy risk under sequential execution. Motivated by this analysis, we propose a privacy-regularized training framework that directly constrains information flow between agent outputs and agent-local sensitive variables. We evaluate our approach across sequential agent pipelines of varying depth on three benchmark datasets, demonstrating stable optimization dynamics and consistent, interpretable privacy-utility trade-offs. Our results show that privacy in agentic LLM systems cannot be guaranteed by local constraints alone and must instead be treated as a system-level property during both training and deployment.

1 INTRODUCTION

Large language models (LLMs) are increasingly deployed as *agentic systems*, where complex user requests are handled not by a single monolithic model, but by multiple interacting agents with specialized roles. Such multi-agent architectures enable modular reasoning, tool use, delegation, and specialization, and are now widely adopted in applications spanning medical decision support, financial analysis, and enterprise automation (Pan et al., 2025; Xiao et al., 2024; Marreed et al., 2025).

In such multi-agent systems, agents can interact in several structural modes, including *parallel* collaboration, where agents operate independently and aggregate their results; *hierarchical* orchestration, where controllers delegate subtasks to workers; and *sequential* pipelines, where agents process information stage-by-stage and pass intermediate representations downstream. Each interaction pattern induces distinct failure modes and security risks. In this work, we focus specifically on *sequential multi-agent pipelines*, a common deployment pattern in which intermediate reasoning states, summaries, or tool outputs produced by one agent serve as input to the next.

Despite their practical advantages, sequential agent pipelines introduce *privacy risks* that are not captured by traditional single-model threat models. In realistic deployments, individual agents often operate with *agent-local sensitive context*, such as proprietary policies, internal routing logic, private memory states, or domain-specific confidential data. Although downstream agents may not be granted direct access to this sensitive information, information about agent-local variables can persist in the representations passed across agent boundaries. Consequently, even when each agent satisfies a local privacy constraint, the sequential system may still exhibit substantial privacy leakage at its final output. Prior research on privacy in LLMs has primarily focused on *single-model* settings, including memorization of training data (Xu et al., 2025; Nasr et al., 2025) and membership inference attacks (Meeus et al., 2025; Mangaokar et al., 2025). While these works provide important model-level insights and privacy guarantees, they do not characterize the privacy risks that emerge *during deployment* when multiple LLM agents exchange and transform information across a pipeline.

More recent work has begun to study privacy and security risks in multi-agent LLM systems, including compositional leakage, adversarial prompt propagation, trust miscalibration, and coordination failures (Patil et al., 2025; Zhang et al., 2025b; Yang et al., 2025; Gomaa et al., 2025) (See Appendix B for broader discussion).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

In particular, Patil et al. (Patil et al., 2025) demonstrate empirically that individually innocuous disclosures from multiple agents can be composed to infer sensitive information, and study inference-time defenses against such attacks. Their formulation, however, focuses on aggregating partial disclosures across collaborating agents rather than characterizing how sensitive information is transformed and propagated through a *feed-forward sequence of learned representations*. An open question is therefore how leakage introduced at individual agent boundaries contributes to the final disclosure in a depth-N sequential pipeline, and whether this contribution depends systematically on the position of the agent within the pipeline.

This distinction is important because, in a sequential agent pipeline, each agent consumes an upstream representation, combines it with local information, and passes a transformed representation to subsequent agents. Intermediate representations, summaries, and tool outputs can therefore preserve statistical dependence on sensitive variables across multiple stages. Consequently, privacy risk depends not only on the amount of sensitive information present locally, but also on *where in the pipeline* that information is introduced and how it propagates through downstream transformations. Existing empirical analyses establish that cross-agent composition can expose sensitive information, but do not provide a depth-dependent characterization relating local representation leakage to final-output leakage, or a training-time mechanism for controlling these intermediate information flows.

We address this gap from an *information-theoretic perspective*. Rather than treating compositional leakage itself as a new phenomenon, we formalize privacy propagation in sequential LLM-agent pipelines, derive a depth-dependent relationship between local representation leakage and final-output leakage, and use this analysis to motivate representation-level privacy control during training. This formulation enables us to study how agent position affects privacy contribution, why leakage introduced earlier in the pipeline can exert greater downstream influence, and how local regularization can be allocated across agents to control global leakage while preserving utility.

Contributions. Our main contributions are summarized as follows:

- We formalize privacy leakage propagation in *sequential* multi-agent LLM pipelines, relating local representation-level leakage to privacy leakage at the final system output.
- We derive a depth-dependent compositional bound under Markov assumptions, characterizing how the contribution of local leakage depends on an agent's position within the sequential pipeline.
- We propose a privacy-regularized training framework that directly controls mutual information (MI) between intermediate agent representations and agent-local sensitive variables, enabling selective and adaptive allocation of privacy regularization across pipeline stages.
- We provide extensive empirical evaluation across medical, financial, and action-based privacy benchmarks, demonstrating consistent and interpretable privacy-utility trade-offs.

2 Problem Setup

We now formalize privacy leakage in sequential multi-agent LLM systems from an information-theoretic perspective. Our focus is on understanding how sensitive information held locally by individual agents propagates through intermediate representations and accumulates under sequential composition. We begin by introducing the system model and threat assumptions, then define a notion of compositional privacy leakage that captures system-level exposure arising from multi-agent execution. This formulation provides the foundation for our theoretical analysis of cumulative leakage and motivates the privacy-regularized training framework in subsequent sections.

2.1 Sequential Agent Pipeline

We consider a sequential multi-agent system in which a single user request is processed through a pipeline of N specialized agents, $\{a_1, \ldots, a_N\}$, where each agent a_i is parameterized by θ_i. For notational simplicity, we refer to the parameterized agent directly as a_i throughout the paper. The request is first handled by agent a_1, whose output is passed to a_2, and so on, until agent a_N produces the final system response.

Let $O_0 \triangleq \varnothing$ denote an empty initial input. For each agent a_i, we define:

- D_i as public or task-relevant inputs to agent a_i,
- S_i as *agent-local sensitive information* accessible only to agent a_i and $S_{<i} = \{S_1, \ldots, S_{i-1}\}$,
- O_i as the output representation produced by agent a_i.

Agent execution follows the sequential update rule

$$O_i = A_i(O_{i-1}, D_i, S_i), \qquad i = 1, \ldots, N, \tag{1}$$

where O_{i-1} denotes the output received from the preceding agent. The final system output is given by O_N.

Conditional Independence Structure. We assume a feed-forward sequential pipeline in which downstream agents do not directly access upstream sensitive variables. In particular, for each $i \in \{1, \ldots, N\}$, we assume

$$S_{<i} \to (O_{i-1}, D_i, S_i) \to O_i \to O_{i+1} \to \cdots \to O_N, \tag{2}$$

where $S_{<i} = (S_1, \ldots, S_{i-1})$.

Interpretation. Eq. 2 formalizes the assumption that sensitive information introduced at earlier stages influences downstream behavior only through the intermediate representations that are explicitly passed along the pipeline. Conditioned on the information available to agent a_i—namely the upstream output O_{i-1}, task-relevant inputs D_i, and its own local sensitive context S_i—the agent's output O_i is independent of upstream sensitive variables $S_{<i}$. As a result, any leakage of upstream sensitive information into later outputs must occur indirectly via sequential transformation and reuse of intermediate representations, rather than through direct access to private agent state. Although S_i is never explicitly transmitted, statistical dependencies between S_i and the agent's output O_i can induce privacy leakage. Because intermediate outputs are reused by downstream agents, such leakage may propagate and accumulate through the pipeline via sequential transformations.

To quantify this phenomenon, we adopt an information-theoretic perspective and measure privacy leakage using MI. We define the *global compositional leakage* as

$$\mathcal{L}_{\text{global}} \triangleq I(O_N\,;\, S_1, \ldots, S_N)\,, \tag{3}$$

which captures how much information about all agent-local sensitive variables $\{S_i\}_{i=1}^N$ can be inferred from the final system output O_N. This definition naturally accounts for leakage arising from sequential composition, even when no individual agent directly discloses its sensitive context.

Assumptions for the theoretical analysis. For the composition results in Section 3, we additionally assume that information about an upstream sensitive variable reaches downstream outputs only through the representations transmitted along the pipeline. Equivalently, conditioned on the transmitted representation and the preceding sensitive context, downstream processing introduces no additional access path to the sensitive variable. For the result that relates global leakage to the unconditional local quantities $I(O_i; S_i)$, we additionally assume that the sensitive variables $S_1, \ldots, S_N$ are mutually independent. This assumption is stated explicitly in the corresponding theorem.

2.2 Threat Model

We assume an adversary that observes the final output O_N and seeks to infer information about the sensitive variables $\{S_i\}_{i=1}^N$. The adversary does not have direct access to agent internals, intermediate representations, or private contexts, but can exploit statistical correlations induced by sequential execution and representation passing. This threat model primarily captures an external observer that has access to the final system output but not to agent-local private contexts or intermediate representations. We separately evaluate leakage from intermediate representations as a diagnostic of information propagation within the pipeline.

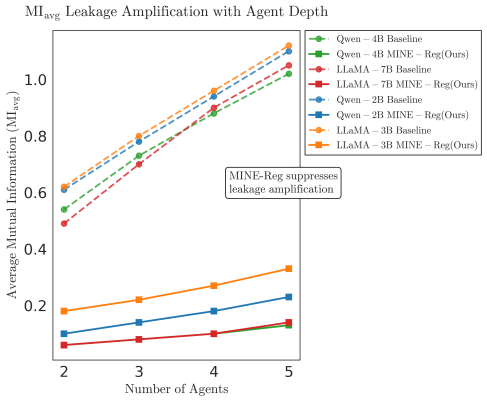


Figure 1: MI$_{\text{avg}}$ leakage as a function of sequential agent depth across different models for MedQA benchmark. Unregularized systems exhibit increasing representation-level MI$_{\text{avg}}$ with pipeline depth, while MI-regularized training substantially flattens this trend across model families.

2.3 Privacy Objective

Our objective is to design learning and execution mechanisms that limit compositional privacy leakage while preserving task utility. Formally, we aim to control the global leakage $\mathcal{L}_{\text{global}}$ subject to maintaining performance on the target task.

Achieving this objective requires understanding how privacy loss arises from the interaction of agents through shared intermediate representations, rather than from any single agent in isolation. In the next section, we formally analyze how privacy leakage behaves under sequential composition and characterize the fundamental limits of agent-local privacy control.

3 Theoretical Analysis for Compositional Privacy Leakage

We now provide a formal analysis of privacy leakage in sequential multi-agent systems. Our goal is to characterize how information about agent-local sensitive variables propagates through intermediate representations and accumulates at the system output.

We show that even when each agent individually satisfies a bounded local privacy constraint, the sequential reuse and transformation of representations can make system-level privacy guarantees substantially weaker than the corresponding local constraints. This theoretical analysis characterizes the gap between agent-local privacy budgets and end-to-end guarantees in composed agent pipelines.

3.1 Per-Agent Leakage

For each agent a_i, we quantify *local leakage* as the MI between the agent's output representation O_i and its agent-local sensitive variable S_i:

$$I(O_i; S_i). \tag{4}$$

A natural design goal is to enforce a per-agent privacy constraint of the form

$$I(O_i; S_i) \leq \tau_i, \tag{5}$$

where $\tau_i \geq 0$ denotes a user-specified leakage threshold or privacy budget for agent a_i.

Such constraints ensure that no single agent output reveals excessive information about its own sensitive context. However, as we show next, local control alone is insufficient to bound the privacy risk of the composed system.

3.2 Sequential Composition and Cumulative Leakage

Our analysis relates the global compositional leakage defined in Eq. 3 to agent-local leakage under sequential execution. Although each agent may satisfy a local information constraint, such constraints alone do not provide a depth-independent guarantee on final-output leakage. We therefore derive a worst-case compositional upper bound that quantifies how local privacy budgets enter the system-level guarantee.

Local vs. Global Leakage. While per-agent leakage bounds defined in Eq. 5 control direct disclosure at individual stages, they do not characterize how information propagates across the pipeline. In sequential settings, correlations introduced upstream are transformed and reused by downstream agents, allowing sensitive dependence to accumulate beyond what any single agent reveals in isolation. As a result, system-level privacy behavior can differ qualitatively from agent-level guarantees.

Theorem 1 (Cumulative Leakage Bound)**.** *Consider a sequential agent pipeline satisfying the Markov structure as given in Eq. 2 and assume that the sensitive variables $\{S_i\}_{i=1}^N$ are mutually independent. If each agent satisfies the local leakage constraint $I(O_i; S_i) \leq \tau_i$, then the global compositional leakage satisfies*

$$I(O_N; S_1, \ldots, S_N) \;\leq\; \sum_{i=1}^{N} 2^{\,N-i}\, \tau_i. \tag{6}$$

Interpretation. Theorem 1 (proof given in Appendix A.1) provides an exponentially weighted *worst-case upper bound* on global leakage in terms of local privacy budgets. Earlier-agent budgets receive larger coefficients because their representations pass through more downstream stages. In the uniform case $\tau_i = \tau$, the bound becomes $(2^N - 1)\tau$. Importantly, this result does not imply that realized leakage grows exponentially with pipeline depth; rather, it characterizes how loose a global guarantee can become when only unconditional local leakage constraints are available. The empirical depth analysis in Section 6.3 separately studies how estimated representation-level leakage changes with pipeline depth.

Position Dependence. The bound assigns larger coefficients to privacy budgets introduced earlier in the pipeline. This motivates examining whether stronger control at earlier stages is useful in practice; we study this empirically through selective regularization in Appendix A.7.

Proposition 1 (Position-Weighted Local Budgets)**.** *Under the conditions of Theorem 1, the coefficient associated with the local privacy budget τ_i in the compositional upper bound is 2^{N-i}. Thus, earlier-stage budgets receive larger worst-case weights than later-stage budgets.*

3.3 Design Implications

Theorem 1 demonstrates that enforcing independent, per-agent privacy constraints does not suffice to control privacy risk in sequential pipelines. Instead, privacy must be treated as a system-level property that explicitly accounts for downstream propagation effects. This observation motivates learning objectives that directly regulate information flow throughout the pipeline, rather than relying on post hoc filtering or isolated local defenses.

In the next section, we introduce a privacy-regularized training framework that operationalizes this principle by explicitly penalizing MI between agent outputs and agent-local sensitive variables during training.

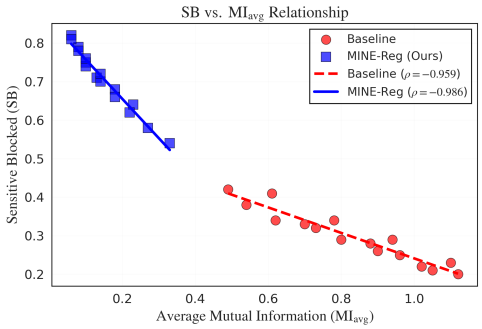


Figure 2: Relationship between $\mathrm{MI}_{\mathrm{avg}}$ and SB for Qwen-4B on FinQA. Lower representation-level MI estimates are strongly associated with higher sensitive blocking under both baseline and MINE-Reg.

4 Privacy-Regularized Sequential Agent Training

We now describe a training framework that explicitly controls privacy leakage in sequential multi-agent systems. Motivated by the theoretical analysis of compositional privacy leakage in Section 3, our approach introduces an information-theoretic regularization term that constrains the dependence between each agent's output and its agent-local sensitive context during training.

4.1 Training Objective

Our objective is to preserve task utility while limiting information leakage at each stage of the agent pipeline. Let y denote the ground-truth task label. The final agent output O_N is used to compute a standard supervised utility loss

$$\mathcal{L}_{\text{utility}} = \ell(O_N, y), \tag{7}$$

where $\ell(\cdot)$ denotes a standard task loss (e.g., cross-entropy).

To control privacy leakage, we penalize the MI between each agent's output and its corresponding sensitive variable:

$$\mathcal{L}_{\text{privacy}} = \sum_{i=1}^{N} \beta_i \, I(O_i; S_i), \tag{8}$$

where $\beta_i \geq 0$ controls the privacy-utility trade-off for agent a_i. The resulting optimization problem is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{utility}} + \mathcal{L}_{\text{privacy}}. \tag{9}$$

The coefficients $\{\beta_i\}_{i=1}^{N}$ can be interpreted as tunable Lagrange multipliers associated with information-flow constraints of the form $I(O_i; S_i) \leq \tau_i$, where τ_i denotes a target leakage threshold. Unless otherwise stated, experiments use fixed β_i values selected through validation tuning. We additionally study adaptive β_i optimization using dual-descent Zhang et al. (2025a) updates to dynamically enforce leakage constraints during training. Additional analysis of fixed, grid-search, and adaptive β_i strategies is provided in Appendix A.7.1.

4.2 Mutual Information Estimation

Direct computation of the MI $I(O_i; S_i)$ is intractable for high-dimensional neural representations produced by large language models. We therefore adopt a variational estimation approach based on the Donsker–Varadhan (DV) representation of the Kullback–Leibler (KL) divergence, as instantiated by Mutual Information Neural Estimation (MINE) (Belghazi et al., 2018).

Donsker-Varadhan Representation. For two distributions P and Q defined on a common support Ω, the KL-divergence admits the variational form

$$D_{\mathrm{KL}}(P \,\|\, Q) = \sup_{T:\Omega\times\Omega\to\mathbb{R}} \left\{ \mathbb{E}_P[T] - \log \mathbb{E}_Q\left[e^T\right] \right\}, \tag{10}$$

where the supremum is taken over all measurable functions T for which the expectations are finite. By setting $P = p(o_i, s_i)$ and $Q = p(o_i)p(s_i)$, the MI between an agent's output O_i and its sensitive variable S_i can be written as

$$I(O_i; S_i) = D_{\mathrm{KL}}\big(p(o_i, s_i) \,\|\, p(o_i)p(s_i)\big) = \sup_{T \in \mathcal{F}} \left\{ \mathbb{E}_{p(o_i,s_i)}[T(o_i, s_i)] - \log \mathbb{E}_{p(o_i)p(s_i)}\left[e^{T(o_i,s_i)}\right] \right\},$$

where $\mathcal{F}$ denotes a parameterized function family.

Neural Estimator Formulation. In practice, we restrict $\mathcal{F}$ to a family of neural networks $T_{\psi_i} : \mathcal{O}_i \times \mathcal{S}_i \to \mathbb{R}$, where $\mathcal{O}_i$ denotes the intermediate representation produced by agent a_i and $\mathcal{S}_i$ denotes the corresponding agent-local sensitive variable, referred to as *critics*, with parameters ψ_i. Given a minibatch of size B, we obtain empirical joint samples $\{(o_i^{(b)}, s_i^{(b)})\}_{b=1}^{B}$ from $p(o_i, s_i)$, and approximate samples from the product of marginals $p(o_i)p(s_i)$ by pairing $o_i^{(b)}$ with shuffled sensitive variables $s_i'^{(b)}$ within the minibatch.

The resulting sample-based MINE estimator is

$$\hat{I}_{\mathrm{MINE}}(O_i; S_i) = \frac{1}{B}\sum_{b=1}^{B} T_{\psi_i}(o_i^{(b)}, s_i^{(b)}) - \log\left(\frac{1}{B}\sum_{b=1}^{B} \exp\left(T_{\psi_i}(o_i^{(b)}, s_i'^{(b)})\right)\right). \tag{11}$$

This estimator provides a differentiable variational lower-bound estimate of the true mutual information and serves as a representation-level leakage diagnostic at agent a_i.

Optimization and Stability. During training, the critic parameters ψ_i are updated to *maximize* $\hat{I}_{\mathrm{MINE}}(O_i; S_i)$ in Eq. 11, while the agent parameters are updated to *minimize* the same quantity as part of the overall privacy-regularized objective. To improve numerical stability of the logarithmic partition term, we maintain an exponential moving average of $\exp(T_{\psi_i})$, following standard and empirically validated MINE implementations (Belghazi et al., 2018). Additional analysis of optimization stability, variance across random seeds, and computational overhead is provided in Appendix A.6.

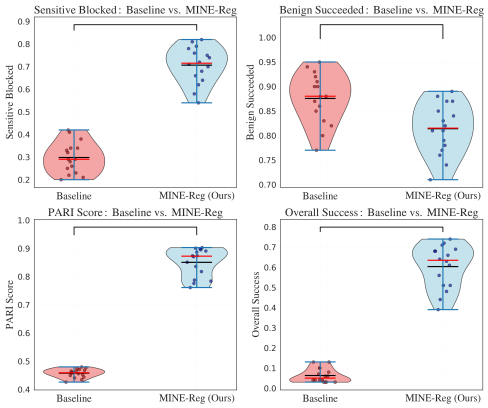


Figure 3: Distributional comparison of baseline and MINE-Reg across privacy and utility metrics for LLaMA-3B. Violin plots show per-run variability with mean (red bar) and dispersion for Sensitive Blocked, Benign Succeeded, PARI score, and Overall Success.

4.3 Optimization Procedure

Training proceeds iteratively and alternates between estimating mutual information and updating the agent parameters to minize $\mathcal{L}_{\text{total}}$ given in Eq. 9. Each iteration consists of three phases: (i) a forward pass through the agent pipeline to obtain $\mathcal{L}_{\text{utility}}$, (ii) updating the MINE critics to obtain accurate estimates of $\hat{I}_{\mathrm{MINE}}(O_i; S_i)$, and (iii) updating the agent parameters by minimizing the combined objective $\mathcal{L}_{\text{total}}$, which balances task utility and privacy regularization. The complete training procedure is given in Algorithm 1 in Appendix A.2.

This optimization enforces an explicit information bottleneck at each agent boundary. As a result, agents learn representations that are sufficient for task completion while minimizing the propagation of agent-local sensitive context through the sequential pipeline.

5 Experimental Setup

We evaluate our privacy-regularized training framework on sequential multi-agent LLM systems across multiple domains. Our experiments are designed to isolate how privacy leakage and task utility scale with agent depth, model capacity, and task domain.

Table 1: MedQA results across LLaMA and Qwen model families (B=Baseline, M=MINE-Reg); Ag = number of agents.

Model	Ag	Meth	CE ↓	MI ↓	SB ↑	BS ↑	PI ↑	PARI ↑
LLaMA-7B	2	B	1.32	0.49	0.42	0.95	0.000	0.480
	2	M	1.47	0.06	0.81	0.89	0.878	0.901
	3	B	1.54	0.70	0.33	0.93	0.000	0.474
	3	M	1.68	0.08	0.78	0.87	0.886	0.899
	4	B	1.73	0.90	0.26	0.90	0.000	0.465
	4	M	1.92	0.10	0.74	0.84	0.889	0.891
	5	B	1.88	1.05	0.21	0.87	0.000	0.456
	5	M	2.06	0.14	0.70	0.81	0.867	0.871
LLaMA-3B	2	B	1.55	0.62	0.34	0.86	0.000	0.453
	2	M	1.72	0.18	0.66	0.78	0.710	0.784
	3	B	1.71	0.80	0.29	0.83	0.000	0.444
	3	M	1.89	0.22	0.62	0.77	0.725	0.788
	4	B	1.88	0.96	0.25	0.80	0.000	0.435
	4	M	2.05	0.27	0.58	0.74	0.719	0.776
	5	B	2.02	1.12	0.20	0.77	0.000	0.426
	5	M	2.21	0.33	0.54	0.71	0.705	0.761

Model	Ag	Meth	CE ↓	MI ↓	SB ↑	BS ↑	PI ↑	PARI ↑
Qwen-4B	2	B	1.40	0.54	0.38	0.94	0.000	0.477
	2	M	1.58	0.06	0.82	0.88	0.889	0.903
	3	B	1.62	0.73	0.32	0.92	0.000	0.471
	3	M	1.77	0.08	0.79	0.85	0.890	0.895
	4	B	1.75	0.88	0.28	0.90	0.000	0.465
	4	M	1.93	0.10	0.75	0.83	0.886	0.887
	5	B	1.89	1.02	0.22	0.88	0.000	0.459
	5	M	2.05	0.13	0.71	0.81	0.873	0.874
Qwen-2B	2	B	1.48	0.61	0.41	0.91	0.000	0.468
	2	M	1.55	0.10	0.76	0.87	0.836	0.874
	3	B	1.66	0.78	0.34	0.88	0.000	0.459
	3	M	1.74	0.14	0.72	0.82	0.821	0.851
	4	B	1.82	0.94	0.29	0.85	0.000	0.450
	4	M	1.93	0.18	0.68	0.79	0.809	0.836
	5	B	1.98	1.10	0.23	0.82	0.000	0.441
	5	M	2.10	0.23	0.64	0.76	0.791	0.818

Benchmarks. We consider three privacy-critical benchmarks: *MedQA* (Jin et al., 2020) for medical reasoning, *FinQA* (Chen et al., 2021) for financial numerical reasoning, and *PrivacyLens* (Shao et al., 2024) for action-based evaluation of contextual privacy norms. Together, these benchmarks cover classification-style reasoning, numerical program execution, and realistic agent trajectories where privacy violations may emerge only through sequential composition.

Sequential Agent Pipelines. We evaluate pipelines of $N \in \{2, 3, 4, 5\}$ agents executed sequentially. Each agent is an independent LLM that communicates exclusively via intermediate representations, reflecting common real-world deployments where reasoning, planning, and policy enforcement are handled by distinct components.

Models. Experiments are conducted using open-weight LLMs from the Qwen (2B, 4B) and LLaMA (3B, 7B, 13B) families,[1] fine-tuned with LoRA adapters (Hu et al., 2021). Agents do not share parameters unless explicitly stated.

Training and Baselines. We compare standard end-to-end training against privacy-regularized training that penalizes MI between intermediate agent representations and agent-local sensitive variables. We additionally compare against adapted privacy-preserving baselines, including DP-style perturbation and privacy regularization methods applied locally at each agent stage. Details of the training objective and sensitive variable definitions are provided in Appendix A.3, while additional baseline implementation details and comparisons are provided in Appendix A.5.

Evaluation. Utility is measured using task-appropriate accuracy or helpfulness metrics, while privacy is evaluated using MI estimates, adversarial leakage accuracy, and composite privacy-utility scores. Formal metric definitions are given in Appendix A.4. We emphasize that $\mathrm{MI}_{\mathrm{avg}}$ is an average of representation-level MINE estimates and is used as an information-flow diagnostic; it is neither an exact estimate nor a certified upper bound on the global leakage $I(O_N; S_{1:N})$. We therefore complement it with a final-output adversarial evaluation that directly matches the threat model.

6 Results

We evaluate MINE-Reg on MedQA (Jin et al., 2020) and FinQA (Chen et al., 2021); PrivacyLens, LLaMA-13B, and extended baseline results are provided in Appendix A.5. All metrics follow Appendix A.4.

6.1 Overall Performance on Medical and Financial Benchmarks

Tables 1 and 2 report MedQA and FinQA results, respectively. Across both domains and all agent depths (2-5 agents), MI-regularized training (MINE-Reg) consistently reduces compositional leakage and improves system-level privacy outcomes, while preserving the majority of benign task performance. Two consistent patterns emerge from the tables: (i) unregularized baselines exhibit progressively worse privacy with increasing depth, and (ii) MINE-Reg flattens this depth-induced

[1]Qwen models: `https://huggingface.co/Qwen/Qwen2-1.5B`, `https://huggingface.co/Qwen/Qwen1.5-4B-Chat`. LLaMA models: `https://huggingface.co/meta-llama/Llama-3.2-3B`, `https://huggingface.co/meta-llama/Llama-2-7b`, `https://huggingface.co/meta-llama/Llama-2-13b`.

Table 2: FinQA results across LLaMA and Qwen model families (B=Baseline, M=MINE-Reg); Ag = number of agents.

Model	Ag	Meth	CE ↓	MI ↓	SB ↑	BS ↑	PI ↑	PARI ↑
LLaMA-7B	2	B	1.38	0.51	0.40	0.94	0.000	0.481
	2	M	1.53	0.06	0.82	0.88	0.882	0.905
	3	B	1.61	0.72	0.31	0.90	0.000	0.472
	3	M	1.78	0.09	0.77	0.84	0.875	0.890
	4	B	1.79	0.93	0.25	0.89	0.000	0.466
	4	M	1.98	0.11	0.73	0.83	0.882	0.892
	5	B	1.95	1.10	0.20	0.86	0.000	0.458
	5	M	2.14	0.15	0.69	0.80	0.864	0.872
LLaMA-3B	2	B	1.62	0.69	0.31	0.85	0.000	0.449
	2	M	1.78	0.21	0.69	0.77	0.696	0.778
	3	B	1.79	0.87	0.26	0.82	0.000	0.441
	3	M	1.96	0.26	0.64	0.75	0.701	0.781
	4	B	1.95	1.02	0.22	0.79	0.000	0.433
	4	M	2.13	0.31	0.60	0.72	0.696	0.768
	5	B	2.10	1.18	0.18	0.76	0.000	0.425
	5	M	2.28	0.37	0.56	0.69	0.686	0.755

Model	Ag	Meth	CE ↓	MI ↓	SB ↑	BS ↑	PI ↑	PARI ↑
Qwen-4B	2	B	1.46	0.57	0.36	0.93	0.000	0.479
	2	M	1.63	0.07	0.80	0.87	0.877	0.900
	3	B	1.69	0.75	0.31	0.91	0.000	0.473
	3	M	1.84	0.09	0.77	0.84	0.880	0.892
	4	B	1.82	0.92	0.27	0.88	0.000	0.464
	4	M	1.99	0.12	0.73	0.81	0.870	0.878
	5	B	1.96	1.08	0.21	0.85	0.000	0.455
	5	M	2.15	0.16	0.69	0.79	0.852	0.863
Qwen-2B	2	B	1.54	0.64	0.34	0.90	0.000	0.468
	2	M	1.62	0.12	0.76	0.85	0.812	0.858
	3	B	1.71	0.82	0.28	0.87	0.000	0.459
	3	M	1.80	0.16	0.72	0.80	0.805	0.842
	4	B	1.88	0.98	0.24	0.84	0.000	0.451
	4	M	1.98	0.21	0.68	0.77	0.786	0.827
	5	B	2.03	1.15	0.19	0.81	0.000	0.442
	5	M	2.15	0.26	0.64	0.74	0.774	0.812

degradation, yielding substantially higher privacy protection at comparable utility. MINE-Reg achieves large reductions in $\mathrm{MI}_{\mathrm{avg}}$ relative to the baseline, typically in the **75–90%** range depending on depth and model scale. On the MedQA benchmark, for Qwen-4B with two agents, $\mathrm{MI}_{\mathrm{avg}}$ drops from 0.54 to 0.06 ($\approx 89\%$ reduction), and remains low even as the pipeline deepens (e.g., 0.13 at five agents versus a baseline of 1.02). Similarly on the FinQA benchmark, LLaMA-7B reduces $\mathrm{MI}_{\mathrm{avg}}$ from 0.51 to 0.06 at two agents ($\approx 88\%$), and from 1.10 to 0.15 at five agents ($\approx 86\%$). These reductions translate into consistent gains in adversarial failure rates: Sensitive Blocked (SB) increases by roughly **+0.35 to +0.50** across most configurations. For example, on MedQA with Qwen-4B, SB improves from $0.38 \rightarrow 0.82$ at two agents and from $0.22 \rightarrow 0.71$ at five agents; on FinQA with LLaMA-7B, SB improves from $0.40 \rightarrow 0.82$ at two agents and from $0.20 \rightarrow 0.69$ at five agents. Notably, the privacy integrity (PI) component mirrors these trends, rising from 0 for baselines (by definition) to ≈ 0.69–0.89 under MINE-Reg, indicating substantial suppression of sensitive dependence.

Utility Preservation. Despite strong privacy improvements, benign accuracy remains comparatively stable under regularization. Across MedQA, Benign Succeeded (BS) typically decreases by **6–8 points** for larger models (e.g., Qwen-4B: $0.94 \rightarrow 0.88$ at two agents; LLaMA-7B: $0.95 \rightarrow 0.89$), and by **8–10 points** for smaller models (e.g., LLaMA-3B: $0.86 \rightarrow 0.78$). On FinQA, benign success is similarly preserved, with LLaMA-7B decreasing from $0.94 \rightarrow 0.88$ at two agents and $0.86 \rightarrow 0.80$ at five agents. Cross-entropy (CE) increases modestly under MINE-Reg (reflecting the expected privacy-utility trade-off), but the degradation remains bounded and does not eliminate the gains in system-level reliability.

Privacy-Aware Reasoning Index (PARI). The composite metric PARI improves substantially across all settings, reflecting simultaneous gains in privacy integrity and preserved reasoning correctness. For Qwen-4B on MedQA, PARI increases from $0.477 \rightarrow 0.903$ at two agents and from $0.459 \rightarrow 0.874$ at five agents; for LLaMA-7B on FinQA, PARI increases from $0.481 \rightarrow 0.905$ at two agents and from $0.458 \rightarrow 0.872$ at five agents. These improvements indicate that MI-regularization yields a consistently better operating point for privacy-sensitive sequential deployments, particularly as agent depth increases.

6.2 Privacy-Utility Trade-off Analysis

Figure 3 provides a distributional comparison between unregularized baselines and MINE-Reg for Llama-3B across varying agent depths on the MedQA. Rather than emphasizing individual configurations, this analysis highlights system-level trends aggregated across experimental settings.

Privacy Gains. The top-left panel shows the distribution of SB. Baseline systems cluster in the low-privacy regime (SB ≈ 0.20–0.40), indicating widespread vulnerability under sequential composition. In contrast, MINE-Reg shifts the distribution upward, with median SB near 0.70 and upper quartiles exceeding 0.80, demonstrating consistent privacy improvements across models and depths.

Utility Preservation. The top-right panel reports BS. While MINE-Reg introduces a modest utility reduction, performance remains concentrated in the high-accuracy regime (BS ≈ 0.78–0.88). The

Table 3: Final-output adversarial inference on MedQA using Qwen-4B. The attacker observes only O_N. Lower attack performance is better.

Agents	Method	Macro Acc.↓	Macro F1↓	BS↑
3	Baseline	0.72	0.70	0.92
3	MINE-Reg	0.30	0.29	0.85
5	Baseline	0.78	0.76	0.88
5	MINE-Reg	0.32	0.31	0.81

substantial overlap with baseline distributions confirms that privacy gains are not achieved via catastrophic utility loss, consistent with the 5–9 point BS drops observed in Tables 1 and 2.

Composite System Performance. The bottom-left panel reports the PARI. Baseline systems cluster around 0.44–0.48, reflecting persistent leakage despite strong benign accuracy. MINE-Reg shifts PARI sharply upward, with medians near 0.87–0.90, indicating a strictly better privacy-utility operating point.

Overall Success (OS). Similarly, the bottom-right panel shows OS, which requires both correct benign behavior and successful privacy protection. Baseline systems fail almost universally (OS < 0.10), whereas MINE-Reg concentrates mass between 0.55 and 0.70. This sharp separation indicates that privacy regularization not only mitigates leakage but also enables reliable end-to-end task completion under safety constraints.

6.3 Leakage Trends with Agent Depth

Figure 1 shows that baseline $\mathrm{MI}_{\mathrm{avg}}$ increases monotonically from two to five agents, whereas MINE-Reg substantially flattens this trend across model families. For example, LLaMA-7B on MedQA increases from 0.49 to 1.05 without regularization, while MINE-Reg remains below approximately 0.35 even at five agents. This empirical trend motivates composition-aware privacy control but should not be interpreted as verification of the worst-case scaling in Theorem 1.

6.4 Information-Theoretic Validation

Figure 2 shows a strong negative association between $\mathrm{MI}_{\mathrm{avg}}$ and SB for Qwen-4B on FinQA ($\rho = -0.959$ baseline; $\rho = -0.986$ MINE-Reg). Lower representation-level MI estimates are therefore consistently associated with stronger sensitive blocking, supporting $\mathrm{MI}_{\mathrm{avg}}$ as a leakage diagnostic. This empirical relationship complements, but does not directly verify, the compositional bound in Section 3.

6.5 Final-Output Adversarial Inference

To directly evaluate the external-observer threat model in Section 2.2, we train an independent attacker that observes only the final system output O_N and predicts the agent-local sensitive variables. Agent parameters are frozen, and the attacker is trained on a disjoint attack-training split.

As shown in table 3 MINE-Reg substantially reduces sensitive-variable recovery from the final output: macro accuracy decreases from 0.72 to 0.30 for three agents and from 0.78 to 0.32 for five agents, while retaining most benign task performance. Because the four-class chance level is 0.25, attack performance remains above chance after regularization, indicating a substantial reduction with residual leakage rather than complete privacy. Adversarial accuracy is an operational attack metric and should not be interpreted as an exact estimate of $I(O_N; S_{1:N})$.

7 Conclusion and Future Work

We analyzed privacy leakage in sequential multi-agent LLM pipelines and studied how agent-local leakage relates to system-level privacy risk under sequential composition. To address this, we introduced MI-regularized training (MINE-Reg), a system-level approach that directly constrains information flow at intermediate representations. The theoretical analysis provides a worst-case compositional characterization, while experiments show that MINE-Reg reduces representation-level leakage and final-output adversarial inference while largely preserving task utility. Several directions remain open for future work. Extending the framework to dynamic and graph-structured agent topologies would better reflect realistic deployments. Exploring adaptive communication and leakage-aware routing strategies as well as scaling the framework to heterogeneous and multimodal agent settings remains an important direction for practical privacy-sensitive deployments.

AI USE STATEMENT

Claude Opus 4.8 (Anthropic) and GPT-5.6 Sol (OpenAI) were used for editing, proofreading, resolving LaTeX formatting issues, and to assist with code development for the implementation of the proposed methods. All AI-assisted text and code were reviewed, verified, and, where necessary, modified by the authors.

REFERENCES

Eugene Bagdasarian, Ren Yi, Sahra Ghalebikesabi, Peter Kairouz, Marco Gruteser, Sewoong Oh, Borja Balle, and Daniel Ramage. Airgapagent: Protecting privacy-conscious conversational agents. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS 2024)*, pp. 3868–3882. ACM, 2024.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, volume 80, Stockholm, Sweden, 2018. PMLR.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914. IEEE, 2022.

Zhaorun Chen, Zhen Xiang, Chaowei Xiao, Dawn Song, and Bo Li. Agentpoison: Red-teaming llm agents via poisoning memory or knowledge bases. *arXiv preprint arXiv:2407.12784*, 2024. URL https://arxiv.org/pdf/2407.12784. Affiliations: University of Chicago; University of Illinois Urbana-Champaign; University of Wisconsin–Madison; University of California, Berkeley.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. Finqa: A dataset of numerical reasoning over financial data. *arXiv preprint arXiv:2109.00122*, 2021. URL https://arxiv.org/pdf/2109.00122. Affiliations: University of California, Santa Barbara; J.P. Morgan; Pennsylvania State University; Carnegie Mellon University.

Amr Gomaa, Ahmed Salem, and Sahar Abdelnabi. Converse: Benchmarking contextual safety in agent-to-agent conversations. *arXiv preprint arXiv:2511.05359*, 2025. URL https://arxiv.org/abs/2511.05359. Affiliations: German Research Center for Artificial Intelligence (DFKI); Microsoft; ELLIS Institute Tübingen and MPI for Intelligent Systems; Tübingen AI Center.

Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. URL https://arxiv.org/pdf/2106.09685. Affiliation: Microsoft Corporation.

Daphne Ippolito, Florian Tramèr, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher Choquette-Choo, and Nicholas Carlini. Preventing generation of verbatim memorization in language models gives a false sense of privacy. In *Proceedings of the 16th International Natural Language Generation Conference*, pp. 28–53, 2023.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*, 2020. URL https://arxiv.org/pdf/2009.13081. Affiliations: MIT, USA; Tongji Medical College, HUST, PRC.

Haris Khan and Sadia Asif. Generative ai agents for controllable and protected content creation. In *NeurIPS 2025 Workshop on Generative AI for Provenance, Copyright, and Control (GenProCC)*, 2025. Published: 24 Sept 2025; Last modified: 07 Nov 2025.

Bogdan Kulynych, Juan Felipe Gomez, Georgios Kaissis, Jamie Hayes, Borja Balle, Flavio P. Calmon, and Jean Louis Raisaro. Unifying re-identification, attribute inference, and data reconstruction risks in differential privacy. In *39th Conference on Neural Information Processing Systems*

(NeurIPS 2025), 2025. Affiliations: Lausanne University Hospital, University of Lausanne, Harvard University, Google DeepMind.

Wenkai Li, Liwen Sun, Zhenxiang Guan, Xuhui Zhou, and Maarten Sap. 1-2-3 check: Enhancing contextual privacy in LLM via multi-agent reasoning. In Leon Derczynski, Jekaterina Novikova, and Muhao Chen (eds.), *Proceedings of the The First Workshop on LLM Security (LLMSEC)*, pp. 115–128, Vienna, Austria, August 2025. Association for Computational Linguistics. ISBN 979-8-89176-279-4. URL https://aclanthology.org/2025.llmsec-1.9/.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations (ICLR 2019)*, 2019. Affiliation: University of Freiburg, Germany.

Neal Mangaokar, Ashish Hooda, Zhuohang Li, Bradley A. Malin, Kassem Fawaz, Somesh Jha, Atul Prakash, and Amrita Roy Chowdhury. What really is a member? discrediting membership inference via poisoning. In *39th Conference on Neural Information Processing Systems (NeurIPS 2025)*, 2025. Affiliations: University of Michigan, Ann Arbor; University of Wisconsin-Madison; Vanderbilt University.

Sami Marreed, Alon Oved, Avi Yaeli, Segev Shlomov, Ido Levy, Aviad Sela, Asaf Adi, and Nir Mashkif. Towards enterprise-ready computer using generalist agent. *arXiv preprint arXiv:2503.01861*, 2025. URL https://arxiv.org/pdf/2503.01861v1.

M. Meeus, I. Shilov, S. Jain, M. Faysse, M. Rei, and Y.-A. de Montjoye. Sok: Membership inference attacks on llms are rushing nowhere (and how to fix it). In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 385–401, Copenhagen, Denmark, 2025. IEEE. doi: 10.1109/SaTML64287.2025.00028.

Fatemehsadat Mireshghallah, Huseyin Inan, Marcello Hasegawa, Victor Rühle, Taylor Berg-Kirkpatrick, and Robert Sim. Privacy regularization: Joint privacy-utility optimization in LanguageModels. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3799–3807, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.298. URL https://aclanthology.org/2021.naacl-main.298/.

Milad Nasr, Javier Rando, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from aligned, production language models. In *The Thirteenth International Conference on Learning Representations (ICLR 2025)*, 2025. URL https://arxiv.org/pdf/2311.17035.

Jiayi Pan, Xingyao Wang, Graham Neubig, Navdeep Jaitly, Heng Ji, Alane Suhr, and Yizhe Zhang. Training software engineering agents and verifiers with swe-gym. In *Proceedings of the 42nd International Conference on Machine Learning (ICML 2025)*, pp. PMLR 267, Vancouver, Canada, 2025. Proceedings of Machine Learning Research.

Vaidehi Patil, Elias Stengel-Eskin, and Mohit Bansal. The sum leaks more than its parts: Compositional privacy risks and mitigations in multi-agent collaboration, 2025. URL https://arxiv.org/abs/2509.14284.

Akshara Prabhakar, Zuxin Liu, Ming Zhu, Jianguo Zhang, Tulika Awalgaonkar, Shiyu Wang, Zhiwei Liu, Haolin Chen, Thai Hoang, Juan Carlos Niebles, Shelby Heinecke, Weiran Yao, Huan Wang, Silvio Savarese, and Caiming Xiong. Apigen-mt: Agentic pipeline for multi-turn data generation via simulated agent-human interplay. *39th Conference on Neural Information Processing Systems (NeurIPS 2025)*, 2025. Affiliation: Salesforce AI Research, USA.

Shaina Raza, Ranjan Sapkota, Manoj Karkee, and Christos Emmanouilidis. Responsible agentic reasoning and ai agents: A critical survey. *TechRxiv preprint*, November 2025. doi: 10.36227/techrxiv.175735299.97215847. URL https://www.techrxiv.org/doi/full/10.36227/techrxiv.175735299.97215847. Affiliations: Vector Institute; Cornell University; University of Groningen.

J. Rosser and Jakob Foerster. Agentbreeder: Mitigating the ai safety risks of multi-agent scaffolds via self-improvement. In *39th Conference on Neural Information Processing Systems (NeurIPS 2025)*, 2025. Affiliations: University of Oxford; FLAIR, University of Oxford.

Yijia Shao, Tianshi Li, Weiyan Shi, Yanchen Liu, and Diyi Yang. Privacylens: Evaluating privacy norm awareness of language models in action. In *38th Conference on Neural Information Processing Systems (NeurIPS 2024), Track on Datasets and Benchmarks*, 2024. Affiliations: Stanford University; Northeastern University; Harvard University.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18. IEEE, 2017.

Saksham Sahai Srivastava and Haoyu He. Memorygraft: Persistent compromise of llm agents via poisoned experience retrieval. *arXiv preprint arXiv:2512.16962*, 2025. URL https://arxiv.org/pdf/2512.16962. Affiliation: School of Computing, University of Georgia, Athens, GA, USA.

Robin Staab, Mark Vero, Mislav Balunovic, and Martin Vechev. Large language models are advanced anonymizers. *arXiv preprint arXiv:2402.13846*, 2024. URL https://arxiv.org/abs/2402.13846.

Yashar Talebirad and Amirhossein Nadiri. Multi-agent collaboration: Harnessing the power of intelligent llm agents. *arXiv preprint arXiv:2306.03314*, 2023. URL https://arxiv.org/abs/2306.03314.

Guanlong Wu, Zheng Zhang, Yao Zhang, Weili Wang, Jianyu Niu, Ye Wu, and Yinqian Zhang. I know what you asked: Prompt leakage via kv-cache sharing in multi-tenant llm serving. In *Network and Distributed System Security (NDSS) Symposium 2025*, San Diego, CA, USA, 2025. doi: 10.14722/ndss.2025.241772. URL https://dx.doi.org/10.14722/ndss.2025.241772.

Yijia Xiao, Edward Sun, Di Luo, and Wei Wang. Tradingagents: Multi-agents llm financial trading framework. In *arXiv preprint arXiv:2412.20138*, 2024. Available at https://arxiv.org/abs/2412.20138.

Yixuan Xu, Antoine Bosselut, and Imanol Schlag. Positional fragility in llms: How offset effects reshape our understanding of memorization risks. In *Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS 2025)*, ETH Zürich, 2025. NeurIPS.

Yingxuan Yang, Huacan Chai, Shuai Shao, Yuanyi Song, Siyuan Qi, Renting Rui, and Weinan Zhang. Agentnet: Decentralized evolutionary coordination for llm-based multi-agent systems. In *39th Conference on Neural Information Processing Systems (NeurIPS 2025)*, 2025. Affiliations: Shanghai Jiao Tong University; Shanghai Innovation Institute.

Botong Zhang, Shuo Li, Ignacio Hounie, Osbert Bastani, Dongsheng Ding, and Alejandro Ribeiro. Alignment of large language models with constrained learning, 2025a. URL https://arxiv.org/abs/2505.19387.

Qixin Zhang, Yan Sun, Can Jin, Xikun Zhang, Yao Shu, Puning Zhao, Li Shen, and Dacheng Tao. Effective policy learning for multi-agent online coordination beyond submodular objectives. In *39th Conference on Neural Information Processing Systems (NeurIPS 2025)*, 2025b. Affiliations: Nanyang Technological University, Singapore; The University of Sydney; Hong Kong University of Science and Technology (Guangzhou); Rutgers University; Sun Yat-sen University.

Yanzhe Zhang and Diyi Yang. Searching for privacy risks in llm agents via simulation. *arXiv preprint arXiv:2508.10880*, 2025. URL https://arxiv.org/abs/2508.10880.

A Appendix

A.1 Proof of Theorem 1

We provide the full derivation of the cumulative leakage bound stated in Theorem 1. The proof relies on repeated applications of the chain rule for mutual information (MI) and the conditional data processing inequality as given below.

Model Setup. We consider the sequential multi-agent system defined in Section 3.1 and briefly recall the notation and assumptions required for the theoretical analysis.

Let S_i denote the agent-local sensitive variable accessible only to agent a_i, O_i the intermediate output produced by agent a_i, and D_i the task-relevant or public input available at stage i. Agent outputs satisfy the update rule $O_i = A_i(O_{i-1}, D_i, S_i)$, with $O_0 = \varnothing$.

We assume that the sensitive variables $\{S_i\}_{i=1}^N$ are mutually independent and that the sequential pipeline satisfies the conditional Markov structure defined in Eq. 2. Finally, each agent satisfies a local information leakage constraint as

$$I(O_i; S_i) \leq \tau_i. \tag{12}$$

Objective. Our goal is to upper bound the total compositional leakage:

$$I(O_N; S_1, \ldots, S_N), \tag{13}$$

in terms of the per-agent leakage budgets $\{\tau_i\}_{i=1}^N$.

Chain Rule Expansion. By the chain rule for MI:

$$I(O_N; S_1, \ldots, S_N) = \sum_{i=1}^{N} I(O_N; S_i \mid S_{<i}). \tag{14}$$

Conditional Data Processing.

Definition 1 (Conditional Data Processing Inequality)**.** *If $X \to Y \to Z$ forms a Markov chain conditioned on W, then*

$$I(X; Z \mid W) \leq I(X; Y \mid W).$$

Applying Definition 1 to the conditional Markov chain in Eq. 2, we obtain:

$$I(O_N; S_i \mid S_{<i}) \leq I(O_i; S_i \mid S_{<i}). \tag{15}$$

Substituting into Eq. 14 yields:

$$I(O_N; S_1, \ldots, S_N) \leq \sum_{i=1}^{N} I(O_i; S_i \mid S_{<i}). \tag{16}$$

Conditional Leakage Decomposition. We define:

$$\mathcal{L}_i := I(O_i; S_i \mid S_{<i}). \tag{17}$$

Since S_i is independent of $S_{<i}$, we have:

$$\begin{aligned} \mathcal{L}_i &= H(S_i \mid S_{<i}) - H(S_i \mid O_i, S_{<i}) \\ &= H(S_i) - H(S_i \mid O_i, S_{<i}). \end{aligned} \tag{18}$$

Similarly,

$$I(O_i; S_i) = H(S_i) - H(S_i \mid O_i). \tag{19}$$

Subtracting the two expressions yields:

$$\begin{aligned} \mathcal{L}_i - I(O_i; S_i) &= H(S_i \mid O_i) - H(S_i \mid O_i, S_{<i}) \\ &= I(S_i; S_{<i} \mid O_i). \end{aligned} \tag{20}$$

Therefore,

$$\mathcal{L}_i = I(O_i; S_i) + I(S_i; S_{<i} \mid O_i). \tag{21}$$

Bounding Cross-Dependencies. [Upstream Leakage Bound] For each i, the following inequality holds:

$$I(S_i; S_{<i} \mid O_i) \leq I(S_{<i}; O_{i-1}).$$

Proof. Since S_i is independent of $S_{<i}$, we have:

$$I(S_i; S_{<i} \mid O_i) = I(S_{<i}; O_i \mid S_i) - I(S_{<i}; O_i).$$

The second term is non-negative, yielding:

$$I(S_i; S_{<i} \mid O_i) \leq I(S_{<i}; O_i \mid S_i).$$

By the conditional Markov structure $S_{<i} \to O_{i-1} \to O_i$ given S_i, the data processing inequality implies:

$$I(S_{<i}; O_i \mid S_i) \leq I(S_{<i}; O_{i-1} \mid S_i) = I(S_{<i}; O_{i-1}),$$

where the final equality follows from independence. □

Recursive Inequality. By Lemma A.1 and Eq. 21:

$$\mathcal{L}_i \leq I(O_i; S_i) + I(S_{<i}; O_{i-1}). \tag{22}$$

Using the chain rule:

$$I(S_{<i}; O_{i-1}) = \sum_{j=1}^{i-1} I(S_j; O_{i-1} \mid S_{<j}). \tag{23}$$

Applying conditional data processing again:

$$I(S_j; O_{i-1} \mid S_{<j}) \leq I(S_j; O_j \mid S_{<j}) = \mathcal{L}_j.$$

Thus,

$$\mathcal{L}_i \leq \tau_i + \sum_{j=1}^{i-1} \mathcal{L}_j. \tag{24}$$

Solving the Recurrence. Define $U_k := \sum_{j=1}^{k} \mathcal{L}_j$. From Eqs. 16 and 17, we note that

$$I(O_N; S_1, \ldots, S_N) \leq U_N. \tag{25}$$

From Eq. 24, we have:

$$U_i = U_{i-1} + \mathcal{L}_i \leq \tau_i + 2U_{i-1}.$$

Solving recursively yields:

$$\begin{aligned} U_1 &\leq \tau_1 \\ U_2 &\leq \tau_2 + 2\tau_1 \\ U_3 &\leq \tau_3 + 2\tau_2 + 4\tau_1 \\ &\vdots \\ U_N &\leq \sum_{i=1}^{N} 2^{N-i}\tau_i. \end{aligned} \tag{26}$$

Conclusion. From Eq. 25, we obtain:

$$\boxed{I(O_N; S_1, \ldots, S_N) \leq \sum_{i=1}^{N} 2^{N-i}\tau_i} \tag{27}$$

In the uniform case $\tau_i = \tau$:

$$I(O_N; S_1, \ldots, S_N) \leq (2^N - 1)\tau.$$

A.2 Training Algorithm

Algorithm 1 Information-Regularized Multi-Agent Training with MINE

Initialize agent parameters $\{\theta_i\}_{i=1}^N$ and estimator parameters $\{\psi_i\}_{i=1}^N$.
for iteration $t = 1$ to T **do**
  Phase 1: Forward Pass
  Compute sequential outputs: $O_1 \to O_2 \to \cdots \to O_N$.
  Compute utility loss: $\mathcal{L}_{\text{utility}} = \ell(O_N, y)$.
  Phase 2: MINE Estimator Update
  for each agent a_i **do**
    Sample joint pairs $(o_i, s_i) \sim p(o_i, s_i)$.
    Sample marginal pairs (o_i, s_i') by shuffling S_i.
    Estimate $\hat{I}_{\text{MINE}}(O_i; S_i)$ using Eq. 11.
    Update $\psi_i \leftarrow \psi_i + \eta_\psi \nabla_{\psi_i} \hat{I}_{\text{MINE}}(O_i; S_i)$.
  end for
  Phase 3: Agent Update
  Compute total loss:
  $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{utility}} + \sum_i \beta_i \hat{I}_{\text{MINE}}(O_i; S_i)$.
  Update agent parameters: $\theta_i \leftarrow \theta_i - \eta_\theta \nabla_{\theta_i} \mathcal{L}_{\text{total}}$.
end for

A.3 Experimental Setup Details

A.3.1 Tasks and Datasets

MedQA (Medical Reasoning). We use the English subset of MedQA (Jin et al., 2020), consisting of multiple-choice clinical questions that require multi-step medical reasoning. MedQA reflects high-stakes healthcare workflows, where intermediate reasoning states may encode sensitive clinical attributes, implicit diagnoses, or proprietary decision logic.

FinQA (Financial Reasoning). FinQA (Chen et al., 2021) is a financial question-answering benchmark grounded in real-world corporate reports. The task requires numerical reasoning over tabular and textual evidence, making it a privacy-sensitive setting where intermediate representations may encode confidential financial indicators or internal analytical strategies.

PrivacyLens (Contextual Privacy Norms). PrivacyLens (Shao et al., 2024) evaluates privacy norm compliance in action-based agent workflows. Each data point is constructed hierarchically from a privacy-sensitive seed (encoded as a five-tuple specifying data type, data subject, sender, recipient, and transmission principle), which is expanded into a natural-language vignette and an executable agent trajectory. The trajectory consists of a sequence of agent actions and environment observations that simulate a privacy-critical scenario, excluding the final action.

In our experiments, agents consume the seed, vignette, and observed trajectory and must generate the final action. PrivacyLens evaluates whether this action leaks sensitive information specified in the seed while maintaining task helpfulness, enabling realistic assessment of sequential privacy leakage.

A.3.2 Sequential Agent Architectures

We evaluate sequential pipelines with $N \in \{2, 3, 4, 5\}$ agents. Each agent is an independent LLM with no shared parameters and communicates exclusively through intermediate representations produced by upstream agents.

The agent roles are instantiated as follows:

- **Agent a_1** (*Input Interpreter*): processes the user input (question or trajectory context) and produces an initial structured representation.
- **Agent a_2** (*Task Reasoner*): consumes upstream representations and performs core task-level reasoning or decision-making.
- **Agent a_3** (*Context Refiner*, when $N \geq 3$): refines and contextualizes intermediate outputs, potentially aggregating partial evidence or reformulating reasoning traces.
- **Agent a_4** (*Privacy-Aware Rewriter*, when $N \geq 4$): performs representation-level rewriting or filtering to reduce leakage of agent-local sensitive information while preserving task utility.

- **Agent a_5** (*Compliance Checker*, when $N = 5$): performs a final consistency and safety pass, ensuring that the output satisfies both task correctness and privacy constraints before release.

This design reflects realistic multi-agent deployments, where interpretation, reasoning, refinement, and policy enforcement are handled by specialized components operating sequentially. Importantly, increasing agent depth introduces additional opportunities for compositional information leakage, making this setting well-suited for studying privacy amplification effects.

A.3.3 MODELS

We evaluate two open-weight model families:

- **Qwen**[2]: 2B and 4B parameter variants
- **LLaMA**[3]: 3B, 7B and 13B parameter variants

All agents are fine-tuned using LoRA (Hu et al., 2021) adapters with rank 8. Unless stated otherwise, agents do not share LoRA parameters.

A.3.4 SENSITIVE VARIABLES

Each agent a_i is associated with an agent-local sensitive variable S_i, which should not be inferable from its output O_i.

- For MedQA and FinQA, S_i includes embeddings of private prompts, internal routing states, or agent-specific policy indicators.
- For PrivacyLens, S_i corresponds to privacy-sensitive seed elements, including data subject, recipient, and transmission principle.

A.3.5 TRAINING OBJECTIVE

The privacy-regularized objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{utility}} + \sum_{i=1}^{N} \beta_i \, \hat{I}_{\text{MINE}}(O_i; S_i),$$

where $\hat{I}_{\text{MINE}}(O_i; S_i)$ is estimated using stabilized variational MI estimator (MINE). (Belghazi et al., 2018) Privacy weights β_i control the privacy-utility trade-off and are ablated in Appendix A.7.

A.3.6 IMPLEMENTATION DETAILS

All models are trained using AdamW Loshchilov & Hutter (2019) with learning rate 2×10^{-5}, weight decay 0.01, and batch size 16. Training uses mixed-precision (bfloat16) computation with gradient clipping norm 1.0. All experiments are trained for 10 epochs using cosine learning-rate decay with 5% warmup steps. LoRA adapters use rank 8 and scaling factor 16. Unless otherwise stated, agents do not share LoRA parameters. MI critics are implemented as two-layer MLPs with hidden dimension 256 and ReLU activations. Critics are optimized using Adam with learning rate 1×10^{-4}. To improve numerical stability of the MINE estimator, we use exponential moving average stabilization of the partition term with decay coefficient 0.99. All experiments are conducted using three random seeds, with reported values averaged across runs. Experiments are executed on NVIDIA H100 GPUs.

A.4 DETAILED EVALUATION METRICS

This section provides formal definitions and implementation details for all privacy and utility metrics used throughout the paper. Our metrics are designed to jointly capture representation-level information leakage, adversarial inference success, and task utility in sequential multi-agent pipelines.

A.4.1 UTILITY METRICS

Cross-Entropy Loss (CE). For MedQA and FinQA, utility is primarily measured using the standard CE loss between the model prediction $\hat{y}$ and the ground-truth label y, where x denotes the input query and $p_\theta(\cdot \mid x)$ is the predictive distribution of the final agent parameterized by θ:

$$\text{CE} = -\mathbb{E}_{(x,y)}\big[\log p_\theta(\hat{y} = y \mid x)\big].$$

[2] https://huggingface.co/Qwen/Qwen2-1.5B, https://huggingface.co/Qwen/Qwen1.5-4B-Chat

[3] https://huggingface.co/meta-llama/Llama-3.2-3B, https://huggingface.co/meta-llama/Llama-2-7b, https://huggingface.co/meta-llama/Llama-2-13b

Lower values indicate better task performance.

Benign Succeeded (BS). BS measures the fraction of benign (non-sensitive) inputs for which the system produces a correct or helpful output, where N_{benign} denotes the number of benign evaluation instances and $\mathbb{I}[\cdot]$ is the indicator function:

$$\text{BS} = \frac{1}{N_{\text{benign}}} \sum_{i=1}^{N_{\text{benign}}} \mathbb{I}[\hat{y}_i = y_i].$$

For PrivacyLens, $\hat{y}_i = y_i$ is replaced by a helpfulness score ≥ 2 on the 0-3 rubric provided by the benchmark. Higher values are better.

A.4.2 Representation-Level Privacy Metrics

Estimated Mutual Information. For each agent a_i, we estimate the MI between its output representation O_i and its agent-local sensitive variable S_i, where O_i denotes the intermediate representation produced by agent a_i and S_i denotes the corresponding sensitive attribute accessible only at that stage:

$$\hat{I}_{\text{MINE}}(O_i; S_i).$$

MI is estimated using variational MI estimators (MINE) (Belghazi et al., 2018). Lower MI indicates weaker statistical dependence and stronger privacy protection.

Average Mutual Information. To summarize privacy leakage across the depth of a sequential pipeline, we report the average MI

$$\text{MI}_{\text{avg}} = \frac{1}{N} \sum_{i=1}^{N} \hat{I}_{\text{MINE}}(O_i; S_i),$$

where N denotes the number of agents in the pipeline and $\hat{I}_{\text{MINE}}(O_i; S_i)$ is the estimated MI at agent a_i.

A.4.3 Adversarial Privacy Metrics

Leakage Accuracy (LA). LA measures the success rate of an adversary $\mathcal{A}$ that attempts to infer sensitive attributes from agent outputs. Let O_i denote the output representation produced by agent a_i, S_i the corresponding ground-truth sensitive attribute, and $\hat{S}_i = \mathcal{A}(O_i)$ the adversary's prediction obtained by applying $\mathcal{A}$ to the agent output:

$$\text{LA} = \frac{1}{N_{\text{sensitive}}} \sum_{i=1}^{N_{\text{sensitive}}} \mathbb{I}[\hat{S}_i = S_i],$$

where $N_{\text{sensitive}}$ denotes the number of sensitive evaluation instances and $\mathbb{I}[\cdot]$ is the indicator function. Lower values indicate stronger privacy protection.

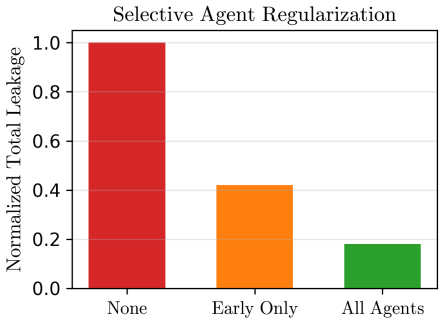


Figure 4: Total compositional leakage under selective agent regularization.

Sensitive Blocked (SB). SB measures the fraction of sensitive cases in which the adversary fails to correctly infer the sensitive attribute:

$$\text{SB} = \frac{1}{N_{\text{sensitive}}} \sum_{i=1}^{N_{\text{sensitive}}} \mathbb{I}[\hat{S}_i \neq S_i] = 1 - \text{LA}.$$

Higher values indicate stronger resistance to adversarial inference.

A.4.4 Joint Privacy–Utility Metrics

Balanced Outcome (BO). BO summarizes the privacy-utility trade-off by averaging benign success and sensitive blocking:

$$\text{BO} = \frac{1}{2}(\text{BS} + \text{SB}).$$

This metric penalizes methods that optimize privacy at the expense of utility, or vice versa, by assigning equal weight to both objectives.

Overall Success (OS). OS measures whether a system simultaneously succeeds on benign inputs and blocks sensitive leakage. For paired benign/sensitive evaluation scenarios, let N_{pairs} denote the number of paired instances, $\text{BenignCorrect}_i \in \{0, 1\}$ indicate whether the benign instance is answered correctly, and $\text{Leakage}_i \in \{0, 1\}$ indicate whether sensitive information is leaked:

$$\text{OS} = \frac{1}{N_{\text{pairs}}} \sum_{i=1}^{N_{\text{pairs}}} \mathbb{I}[\text{BenignCorrect}_i = 1 \ \wedge \ \text{Leakage}_i = 0].$$

A.4.5 PRIVACYLENS-SPECIFIC METRIC MAPPING

Table 4: Additional MedQA results for LLaMA-13B (B = Baseline, M = MINE-Reg).

Agents	Method	$\text{MI}_{\text{avg}} \downarrow$	SB ↑	BS ↑	PARI ↑
2	B	0.46	0.44	0.94	0.480
2	M	0.05	0.83	0.89	0.901
3	B	0.68	0.36	0.92	0.469
3	M	0.07	0.79	0.87	0.887

PrivacyLens evaluates action-based privacy leakage using Leakage Rate (LR) and Adjusted Leakage Rate (LRh), computed over actions with positive helpfulness. We map these quantities to our metrics as follows:

- **Leakage Accuracy (LA)** := LRh
- **Sensitive Blocked (SB)** := $1 - \text{LRh}$
- **Benign Succeeded (BS)** := fraction of actions with helpfulness ≥ 2

This mapping allows consistent comparison between representation-level MI metrics and action-based privacy outcomes.

A.4.6 COMPOSITE METRIC: PRIVACY-AWARE REASONING INDEX (PARI)

To summarize privacy-utility trade-offs in a single scalar metric, analogous to the Responsible Reasoning Index (RRI) introduced in Raza et al. (2025), we define the *Privacy-Aware Reasoning Index (PARI)* as:

$$\text{PARI} = w_1 \cdot RC + w_2 \cdot OT + w_3 \cdot PI + w_4 \cdot RT.$$

PARI is intended as a system-level summary metric that aggregates multiple existing privacy and utility signals, rather than introducing new primitive measurements.

Reasoning Correctness (RC). We define reasoning correctness directly in terms of benign task success using the Benign Succeeded (BS) metric introduced earlier:

$$RC := \text{BS}.$$

This choice avoids introducing a separate correctness score and ensures direct consistency between task-level utility evaluation and the composite metric.

Operational Transparency (OT). Operational Transparency measures the stability of intermediate representations under benign input perturbations. Let q denote a benign input query, $q + \tau$ a semantically equivalent perturbation, and $O(\cdot)$ the final system representation produced by the agent pipeline. We define:

$$OT := 1 - \mathbb{E}\big[\|O(q) - O(q + \tau)\|_2\big].$$

In our experiments, representation stability remains consistent across methods; thus, OT is treated as constant and does not affect relative comparisons.

Privacy Integrity (PI). Privacy Integrity quantifies the relative reduction in representation-level leakage. Let MI_{avg} denote the average estimated MI across agents for a given method, and $\text{MI}_{\text{baseline}}$ the corresponding value for the unregularized baseline. We define:

$$PI := 1 - \frac{\text{MI}_{\text{avg}}}{\text{MI}_{\text{baseline}}}.$$

Table 5: PrivacyLens results for lightweight models (B = Baseline, M = MINE-Reg).

Model	Ag	Meth	MI_{avg} ↓	LA ↓	SB ↑	BS ↑	OS ↑	PARI ↑
Qwen-2B	2	B	0.64	0.61	0.39	0.90	0.12	0.468
Qwen-2B	2	M	0.12	0.25	0.75	0.86	0.65	0.862
Qwen-2B	3	B	0.82	0.68	0.32	0.87	0.07	0.459
Qwen-2B	3	M	0.17	0.29	0.71	0.80	0.57	0.846
Qwen-2B	4	B	0.98	0.73	0.27	0.84	0.04	0.451
Qwen-2B	4	M	0.22	0.34	0.66	0.77	0.51	0.832
LLaMA-3B	2	B	0.69	0.66	0.34	0.86	0.10	0.453
LLaMA-3B	2	M	0.21	0.37	0.63	0.78	0.49	0.781
LLaMA-3B	3	B	0.87	0.71	0.29	0.83	0.06	0.444
LLaMA-3B	3	M	0.26	0.41	0.59	0.75	0.44	0.776
LLaMA-3B	4	B	1.02	0.75	0.25	0.80	0.04	0.435
LLaMA-3B	4	M	0.31	0.45	0.55	0.72	0.40	0.764

Reproducibility and Traceability (RT). Reproducibility and Traceability capture whether system execution is deterministic and auditable. Since all experiments enforce fixed routing, deterministic decoding, and comprehensive logging, this component is fixed:

$$RT := 1.$$

Interpretation and Metric Mapping. PARI does not introduce new evaluation primitives; rather, it aggregates previously defined metrics into a single scalar score. Specifically, RC corresponds directly to BS, PI is a normalized function of average representation-level MI, OT captures representation stability under benign perturbations, and RT encodes deterministic execution. This decomposition allows PARI to summarize the privacy–utility trade-off while preserving interpretability in terms of underlying metrics.

Weights. Unless otherwise stated, we use:

$$(w_1, w_2, w_3, w_4) = (0.3, 0.1, 0.5, 0.1),$$

reflecting the central importance of privacy integrity in our evaluation.

A.5 Additional Experiments

We provide additional validation of our framework on the **PrivacyLens** benchmark (Shao et al., 2024), which is specifically designed to evaluate privacy leakage and adversarial inference risks in language models. Unlike MedQA and FinQA, PrivacyLens does not emphasize task-level reasoning, but instead probes the model's ability to resist sensitive attribute inference under controlled adversarial settings.

To isolate privacy behavior in resource-constrained regimes, we report results only for **Qwen-2B** and **LLaMA-3B**. All evaluation metrics follow the same definitions as in Appendix A.4, and all MI estimates are computed using frozen estimators.

A.5.1 PrivacyLens Results

Discussion. The PrivacyLens results reinforce the core findings of the main paper. Across both Qwen-2B and LLaMA-3B, MI-regularized training reduces $\mathrm{MI}_{\mathrm{avg}}$ by approximately **65–80%** relative to unregularized baselines, leading to substantial improvements in resistance to sensitive attribute inference. SB increases consistently across agent depths, while LA decreases in near-linear correspondence with $\mathrm{MI}_{\mathrm{avg}}$.

As expected, absolute benign performance (BS) is slightly lower than in MedQA and FinQA, reflecting the adversarial nature of PrivacyLens. Nevertheless, utility degradation remains moderate, and the resulting OS improves by more than an order of magnitude in most settings. These results confirm that the benefits of MI-regularization extend beyond task-centric benchmarks to dedicated privacy stress tests.

Importantly, the observed trends mirror those in the main results: unregularized systems exhibit a pronounced increase in leakage with agent depth, while MI-regularization effectively suppresses cumulative leakage. This consistency across benchmarks strengthens the case for MI as a principled, system-level control mechanism for privacy in sequential multi-agent LLM deployments.

A.5.2 Additional Experiments on Larger Models

To evaluate whether the observed privacy-utility trends persist beyond lightweight model families, we additionally conduct experiments using LLaMA-13B on the MedQA benchmark under two-agent and three-agent sequential pipelines.

A.5.3 LLaMA-13B Results

The LLaMA-13B experiments exhibit trends consistent with the main paper. Unregularized pipelines show increasing leakage as agent depth increases, while MINE-Reg substantially suppresses compositional leakage with only modest utility degradation. For example, under the three-agent setting, $\mathrm{MI}_{\mathrm{avg}}$ decreases from 0.68 to 0.07, while benign success remains comparatively stable ($0.92 \rightarrow 0.87$). These results indicate that the effectiveness of MI-regularized training extends to larger-scale model families beyond the lightweight settings explored in the main paper.

A.5.4 Comparison with Alternative Privacy Interventions

Table 6: Optimization stability and computational overhead on MedQA. Lower variance indicates more stable optimization dynamics.

Model	Ag	Metric	Base	Ours
7B	2	Std($\mathrm{MI}_{\mathrm{avg}}$)	0.09	0.02
7B	2	Std(BS)	0.04	0.03
7B	2	Time / iter	1.00×	1.15×
7B	3	Std($\mathrm{MI}_{\mathrm{avg}}$)	0.11	0.03
7B	3	Std(BS)	0.05	0.04
7B	3	Time / iter	1.00×	1.18×
13B	2	Std($\mathrm{MI}_{\mathrm{avg}}$)	0.07	0.02
13B	2	Std(BS)	0.03	0.03
13B	2	Time / iter	1.00×	1.22×
13B	3	Std($\mathrm{MI}_{\mathrm{avg}}$)	0.10	0.03
13B	3	Std(BS)	0.04	0.03
13B	3	Time / iter	1.00×	1.25×

We additionally compare MINE-Reg against representative privacy-preserving interventions adapted to the sequential multi-agent setting. Most existing privacy methods are designed for single-model deployments and do not directly account for representation reuse and sequential dependence across agents. To enable a fair comparison, we consider two adapted baselines:

DP-style Noise. Gaussian noise is injected into each intermediate representation before it is passed to the downstream agent.

Privacy Regularization (PR). Inspired by prior privacy-utility regularization approaches (Mireshghallah et al., 2021), we apply a distribution-flattening regularizer independently at each agent.

In all cases, interventions are applied locally at every stage of the sequential pipeline, and evaluation uses the same metrics as the main experiments. All baselines are implemented under identical model, optimization, and training-budget settings for fair comparison. Results on MedQA are summarized in Table 7.

Across both lightweight and larger-scale models, DP-style perturbation and privacy regularization reduce local leakage but introduce notable utility degradation or weaker compositional control as depth increases. In contrast, MINE-Reg consistently achieves a stronger privacy-utility trade-off by directly regularizing representation-level information flow across the sequential pipeline. For example, under the 3-agent LLaMA-13B setting, MINE-Reg improves SB from 0.36 to 0.79 while preserving substantially higher benign success than DP-style perturbation (0.87 vs. 0.57 BS). These results suggest that local perturbation mechanisms alone are insufficient to control cumulative leakage in sequential multi-agent systems.

A.6 Optimization Stability and Computational Overhead

We further analyze the optimization stability and computational overhead of MINE-Reg on the MedQA benchmark. In our setting, stability is reflected in three aspects: (i) variance of MI estimates across random seeds, (ii) consistency of downstream task performance measured through BS, and (iii) computational overhead introduced by the MINE estimator.

Table 6 summarizes the standard deviation of $\mathrm{MI}_{\mathrm{avg}}$ and BS across multiple runs, together with the relative training time per iteration.

Table 7: Comparison with alternative privacy interventions on MedQA. DP-style noise and PR reduce local leakage but either degrade utility or fail to control compositional leakage effectively as depth increases.

Model	Ag	Method	$\mathrm{MI}_{\mathrm{avg}} \downarrow$	SB ↑	BS ↑	PARI ↑
LLaMA-3B	2	Base	0.69	0.31	0.85	0.449
	2	DP	0.18	0.68	0.52	0.552
	2	PR	0.36	0.57	0.74	0.662
	2	Ours	**0.21**	**0.69**	**0.77**	**0.778**
	3	Base	0.87	0.26	**0.82**	0.441
	3	DP	0.22	0.64	0.48	0.521
	3	PR	0.48	0.52	0.72	0.632
	3	Ours	**0.26**	**0.64**	0.75	**0.781**
LLaMA-13B	2	Base	0.46	0.44	**0.94**	0.480
	2	DP	0.05	0.74	0.61	0.651
	2	PR	0.27	0.66	0.87	0.742
	2	Ours	**0.05**	**0.83**	0.89	**0.901**
	3	Base	0.68	0.36	**0.92**	0.469
	3	DP	0.06	0.70	0.57	0.603
	3	PR	0.38	0.60	0.85	0.721
	3	Ours	**0.07**	**0.79**	0.87	**0.887**

Across all settings, MINE-Reg exhibits substantially lower variance in $\mathrm{MI}_{\mathrm{avg}}$ compared to unregularized baselines, indicating more stable estimation of representation-level leakage during training. At the same time, BS remains consistent across runs, suggesting that privacy regularization does not introduce unstable utility behavior.

The computational overhead introduced by MINE estimation remains moderate, ranging from approximately $1.15\times$ to $1.25\times$ relative training cost per iteration depending on model scale and agent depth. Importantly, the overhead scales predictably with pipeline complexity and does not lead to unstable optimization dynamics in practice.

A.7 Ablation Studies

We conduct a series of ablation studies to analyze the behavior of the proposed privacy-regularized training framework and to validate key design choices. Unless otherwise stated, all ablations are evaluated using validation accuracy as the utility metric and estimated MI leakage $\hat{I}_{\mathrm{MINE}}(O_i; S_i)$ as the privacy metric.

A.7.1 Effect of Privacy Weight and Adaptive β Optimization

We first examine the effect of the privacy regularization strength by sweeping $\beta \in \{0, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$ with uniform allocation $\beta_1 = \beta_2 = \beta$. This experiment isolates the role of β as a global control parameter governing the privacy-utility trade-off.

Figure 5 reveals a smooth and monotonic trade-off between privacy and utility. As β increases, total information leakage decreases consistently, while validation accuracy degrades gradually rather than catastrophically. Notably, the utility curve exhibits diminishing sensitivity beyond moderate leakage reductions, indicating that a substantial fraction of sensitive information can be removed before task-relevant performance is severely affected.

This behavior suggests that β functions as a stable and interpretable privacy knob, enabling practitioners to select operating points that achieve meaningful privacy gains without incurring abrupt performance collapse. The absence of sharp phase transitions further indicates that the optimization landscape remains well-conditioned across a wide range of privacy budgets.

Adaptive β Optimization. In addition to fixed privacy weights, we also study adaptive strategies for optimizing β_i during training. Motivated by constrained optimization methods, we interpret β_i as Lagrange multipliers associated with target leakage constraints $I(O_i; S_i) \leq \tau_i$ and update them dynamically using dual descent:

$$\beta_i \leftarrow \max\Big(0,\ \beta_i + \eta_\beta \big(\hat{I}_{\mathrm{MINE}}(O_i; S_i) - \tau_i\big)\Big). \tag{28}$$

We compare three strategies: (i) fixed β, (ii) grid-search tuning of β_i, and (iii) adaptive dual-descent optimization. Table 8 summarizes results on MedQA using Qwen-4B under two-agent and three-agent settings.

Table 8: Comparison of fixed and adaptive β optimization strategies on MedQA using Qwen-4B.

Agents	Method	$\text{MI}_{\text{avg}} \downarrow$	SB ↑	BS ↑	PARI ↑
2	Fixed β	0.06	0.82	0.88	0.903
2	Grid search β_i	0.06	0.84	0.89	0.915
2	Dual descent	0.04	0.85	0.89	0.922
3	Fixed β	0.08	0.79	0.85	0.895
3	Grid search β_i	0.09	0.81	0.86	0.907
3	Dual descent	0.06	0.83	0.85	0.915

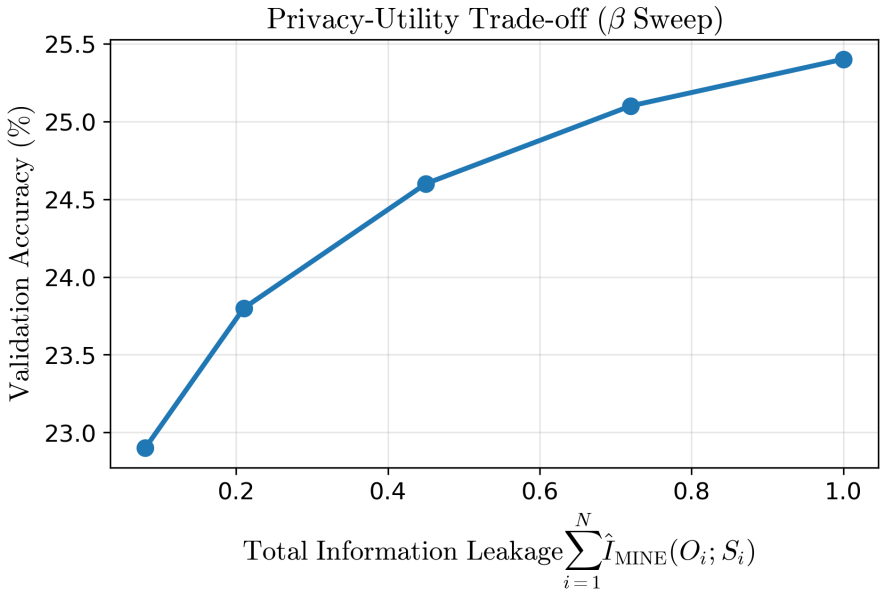


Figure 5: Privacy-utility trade-off as a function of the privacy weight β, measured by validation accuracy versus total information leakage $\sum_{i=1}^{N} \hat{I}_{\text{MINE}}(O_i; S_i)$, for N=3 sequential-agent pipeline using Qwen-2B evaluated on MedQA.

Adaptive dual-descent optimization consistently achieves stronger privacy-utility trade-offs than fixed weighting strategies, reducing leakage while preserving benign task performance. Compared to exhaustive grid search, dual descent achieves comparable or better performance with substantially lower tuning cost.

A.7.2 Selective Agent Regularization

In our experiments, fixed β training requires approximately $1\times$ training cost, grid-search tuning requires $\sim 4\times$, and adaptive dual descent requires approximately $\sim 1.5\times$ training cost. These results suggest that adaptive privacy weighting provides a practical and stable mechanism for enforcing information-flow constraints in sequential multi-agent systems.

We study whether privacy regularization must be applied to all agents or only to a subset of the pipeline. We compare three settings: (i) no agents regularized, (ii) only early agents regularized, and (iii) all agents regularized. All settings use the same total privacy budget.

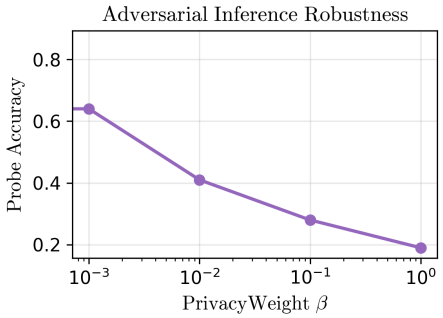


Figure 6: Adversarial probe accuracy for predicting sensitive variables from agent outputs.

Figure 4 shows that regularizing only early agents substantially reduces total leakage relative to the unregularized baseline, highlighting the importance

Table 9: Conditional leakage diagnostics for a five-agent Qwen-4B pipeline on MedQA. Reported values are MINE estimates, not certified information-theoretic bounds.

Method	i=1	i=2	i=3	i=4	i=5	**Sum**
Baseline	0.61	0.72	0.83	0.94	1.06	4.16
MINE-Reg	0.04	0.05	0.07	0.08	0.10	0.34

of controlling information release at early stages. However, significant residual leakage persists due to continued leakage from unregularized downstream agents.

Full regularization across all agents consistently achieves the lowest leakage, demonstrating that privacy risks are distributed throughout the pipeline rather than localized to initial stages. This result emphasizes the need for system-level privacy control and cautions against partial defenses that leave downstream agents unregulated.

A.7.3 Representation Compression Effects

We analyze how MI regularization affects the information content of intermediate representations. Specifically, we measure $I(O_i; y)$ (task relevance) and $I(O_i; S_i)$ (privacy leakage) as β increases.

Figure 7 illustrates that increasing β leads to a pronounced reduction in $I(O_i; S_i)$, while $I(O_i; y)$ decreases much more gradually. This asymmetric compression indicates that the regularizer preferentially suppresses sensitive information rather than indiscriminately collapsing representations.

The preservation of task-relevant signal explains the smooth utility degradation observed in earlier experiments and suggests that the model learns to discard privacy-sensitive features while retaining predictive structure. This behavior aligns with the intended role of information-theoretic regularization as a selective bottleneck rather than a blunt constraint.

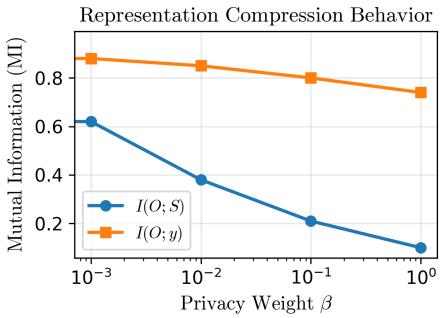


Figure 7: Effect of privacy regularization on task-relevant and sensitive information in intermediate representations.

A.7.4 Robustness to Adversarial Inference

Finally, we evaluate whether reductions in MI translate into practical robustness against inference attacks. We train a lightweight adversarial probe to predict the sensitive attribute S_i from agent outputs O_i under varying privacy budgets.

As shown in Figure 6, adversarial probe accuracy decreases monotonically as β increases, closely tracking the reduction in estimated MI. At higher privacy weights, probe performance approaches chance level, indicating that sensitive attributes are no longer reliably recoverable from intermediate representations.

This result confirms that MI regularization yields operational privacy benefits beyond metric-level improvements, directly limiting the ability of downstream attackers to exploit latent representations for sensitive inference.

A.8 Conditional Leakage Diagnostic

To more closely connect the empirical diagnostics with the global leakage definition, we additionally estimate conditional terms $I(O_i; S_i \mid S_{<i})$ using conditional MINE critics. By the chain rule and the conditional data-processing relation under the assumptions of Section 3,

$$I(O_N; S_{1:N}) \leq \sum_{i=1}^{N} I(O_i; S_i \mid S_{<i}). \tag{29}$$

As shown in table 9 MINE-Reg reduces the estimated conditional leakage at every stage. Since MINE is a variational estimator, the empirical sum is used only as a theoretically motivated diagnostic and is not treated as a certified numerical upper bound on $I(O_N; S_{1:N})$.

B Extended Related Work

B.1 Memorization and Training-Data Privacy

A large body of work on LLM privacy has focused on memorization, training-data extraction, and membership inference. Prior studies show that language models can reproduce rare or repeated training examples and reveal whether particular records were included in the training data (Shokri et al., 2017; Carlini et al., 2022; Ippolito et al., 2023; Xu et al., 2025; Nasr et al., 2025). Differentially private training and related approaches aim to mitigate such record-level leakage (Li et al., 2025; Kulynych et al., 2025). These methods primarily address privacy risks originating from training data, whereas our setting concerns inference-time information flow through intermediate representations in a deployed multi-agent pipeline.

B.2 Inference-Time Privacy and Attribute Inference

More recent work studies privacy leakage that arises during inference, where models can infer sensitive attributes, preferences, or intentions from prompts and contextual information even when those attributes are not explicitly provided (Staab et al., 2024; Wu et al., 2025). Related work on attribute inference and user profiling similarly demonstrates that textual and contextual signals can expose latent private information. These settings primarily consider privacy leakage within a single model or interaction context, while sequential multi-agent systems introduce an additional channel through which sensitive information may be transformed and propagated across multiple stages.

B.3 Contextual Integrity and Privacy Evaluation

Contextual-privacy research views privacy as a property of appropriate information flow rather than secrecy alone (Bagdasarian et al., 2024). Benchmarks such as PrivacyLens operationalize this perspective by evaluating whether model actions violate contextual-integrity constraints in realistic scenarios (Shao et al., 2024). Such evaluations provide an important action-level view of privacy behavior, but primarily assess final decisions or actions rather than the information retained in intermediate representations. Our work complements this perspective by studying how agent-local sensitive information can persist through intermediate outputs and contribute to downstream privacy risk.

B.4 Multi-Agent Systems, Communication, and Memory Risks

LLM-based multi-agent systems increasingly support collaborative reasoning, tool use, delegation, and autonomous workflows (Talebirad & Nadiri, 2023; Khan & Asif, 2025; Prabhakar et al., 2025; Zhang & Yang, 2025). These systems frequently exchange intermediate representations, tool outputs, and memory states, creating privacy and security risks that do not arise in the same form in isolated single-model settings.

Recent studies identify vulnerabilities associated with agent communication, persistent memory, poisoning, and cross-agent inference (Gomaa et al., 2025; Rosser & Foerster, 2025; Chen et al., 2024; Srivastava & He, 2025). Such results highlight that agent topology, communication structure, and memory mechanisms can materially affect system-level risk. Existing defenses often rely on sanitization, prompt filtering, memory isolation, or access control, but provide limited characterization of how sensitive information propagates through a feed-forward sequence of learned agent representations.

B.5 Compositional Privacy Leakage

Recent work has begun to study privacy as a compositional phenomenon in multi-agent systems. In particular, individually limited or seemingly benign disclosures from multiple agents can be combined to infer sensitive information (Patil et al., 2025; Zhang et al., 2025b; Yang et al., 2025; Gomaa et al., 2025). These studies establish that privacy risk may emerge at the system level even when individual components appear locally safe.

Our setting differs in focusing specifically on *sequential* agent pipelines, where each agent transforms an upstream representation before passing it to the next stage. Rather than treating compositional leakage as a new phenomenon, we study how agent-local representation leakage relates to final-output privacy risk under sequential execution. We provide an information-theoretic characterization of this relationship and use it to motivate a training-time regularization strategy that constrains sensitive information at intermediate representation boundaries.

B.6 BROADER IMPACTS

This work studies privacy leakage in sequential multi-agent LLM systems and proposes a privacy-regularized training framework to reduce information exposure across agent interactions. A potential positive impact of this work is improving the safety and privacy of agentic AI systems deployed in domains such as healthcare, finance, and enterprise decision-making, where sensitive information may propagate through intermediate reasoning stages.

At the same time, the techniques and analyses introduced in this paper could potentially be misused to better understand or optimize privacy attacks against multi-agent systems. Our experiments are conducted on benchmark datasets and open-weight models, and we do not release any sensitive user data or high-risk system artifacts.