LEARNING WHAT TO SHARE AND WHAT TO COMPLEMENT FOR MULTI-AGENT COLLABORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-agent collaboration sharing strategies facilitate knowledge exchange across agents, yet struggle to balance privacy protection with effective collaboration when agents hold different private knowledge. We redesign the communication boundaries around these knowledge differences and propose a privacy-aware Multi-agent collaboration framework that uses them as structural signals throughout the reciprocal communication process. Relevance-aware Adaptive Disclosure (RAD) governs outgoing task sharing through a self-calibrated disclosure boundary, moving beyond task fragmentation. Complementary Expertise Integration (CEI) governs returned peer knowledge through geometry-guided complementarity, preserving useful expertise while suppressing redundant reasoning. Across medical and legal tasks on MedQA and LegalQA, our framework consistently improves both open-source and closed-source LLMs, achieving up to 15.3% relative gains in task utility and 9.3% in privacy–utility trade-off over baselines.

1 INTRODUCTION

Independently operated Large Language Model (LLM)-based agents, such as those deployed across hospitals or law firms (Yue et al., 2025; Liu et al., 2025; Wang et al., 2025), often retain private knowledge that is distributed differently across organizations (Zhang et al., 2024; Bonawitz et al., 2021). We refer to such non-uniform private knowledge as *heterogeneous private knowledge* (Shi et al., 2025). This heterogeneity motivates Multi-agent collaboration, as distinct private knowledge gives rise to complementary expertise across agents (Li et al., 2024). Such collaboration typically spans two communication boundaries, where a target agent shares task information with peer agents (task sharing) and receives their outputs in return (output sharing) (Xu et al., 2025; Reid et al., 2026). For example, a cardiology-focused hospital may seek gastroenterology expertise when a patient's symptoms extend beyond its specialty (Tang et al., 2024). In such settings, balancing patient privacy (*e.g.*, identity and medical history) with the diagnostic value of collaboration remains challenging.

Existing methods handle the two communication boundaries with simplified sharing strategies, as shown in Fig. 1. For task sharing, sharing the full task preserves useful context but exposes sensitive patient information (Chen et al., 2026; Li et al., 2026). Fragmentation reduces this exposure, yet may discard cross-symptom context needed for specialist reasoning (Hui et al., 2026). For instance, separating gastrointestinal symptoms from the patient's cardiac history may protect sensitive details but also obscure clinically relevant information. When peer agents return their outputs, final answers or claims convey the conclusion without capturing the expert decision process (Tang et al., 2024; Besrour et al., 2025; Jiang et al., 2026). After all, simply stating that "this is not a stomach problem" is far less useful than explaining why symptoms support that judgment.

These simplified strategies still struggle to balance privacy protection with effective collaboration because they do not fully exploit the heterogeneity which motivates collaboration in the first place. To examine whether such heterogeneity provides a useful coordination signal, we conduct an observational study in Fig. 2. We find that agents with stronger private-memory coverage in a domain make more reliable decisions and reasoning, revealing expertise that is particularly valuable to peers with weaker coverage. *This observation motivates us to redesign both communication boundaries around heterogeneity itself, rather than relying on task fragmentation or compressed output sharing.*

However, this introduces two distinct challenges. When sending task information to peer agents, the goal is no longer simply to conceal sensitive content, but to determine which information remains

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.

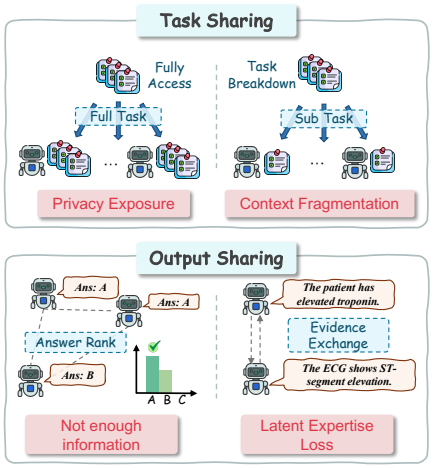


Figure 1: Limitations of task and output sharing paradigms: task sharing may cause privacy exposure or context fragmentation, while output sharing may provide insufficient information or fail to preserve latent expertise.

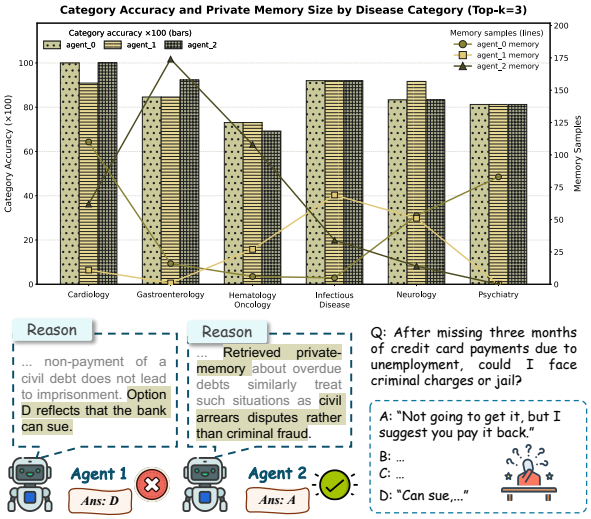


Figure 2: Heterogeneous private knowledge induces complementary agent expertise, as private-memory differences lead agents to retrieve distinct evidence and produce different reasoning outcomes.

valuable under heterogeneous private knowledge, *motivating the claim that disclosure should be selectively conditioned on such knowledge differences.* When receiving outputs from peer agents, richer peer information is not necessarily more useful, since expertise from heterogeneous peers can still overlap with what the target already knows. *We therefore argue that returned knowledge should be selected for its complementary value rather than shared indiscriminately.*

To instantiate these claims, we reconstruct a privacy-aware Multi-agent collaboration framework around heterogeneity-guided communication boundaries. On the outgoing task-sharing boundary, Relevance-aware Adaptive Disclosure (RAD) reframes task sharing as expertise-preserving disclosure and derives a self-calibrated, query-specific disclosure boundary with Bayes-optimal decisions under the local expertise distribution, avoiding fixed global criteria or task fragmentation. On the returning output-sharing boundary, Complementary Expertise Integration (CEI) introduces geometry-guided sharing that identifies peer expertise extending beyond the target agent's existing knowledge while suppressing redundant reasoning. Our main contributions are summarized as follows:

- We identify heterogeneous private knowledge as the key signal for overcoming existing sharing strategies. Based on this insight, we reconstruct privacy-aware Multi-agent collaboration around two heterogeneity-guided decisions on *what to disclose* and *what to absorb*.
- We instantiate these decisions across the reciprocal communication process. RAD derives a self-calibrated, Bayes-optimal disclosure boundary for outgoing task information, while CEI uses geometry-guided complementarity to select useful peer expertise.
- We validate our framework on medical and legal reasoning tasks with both open- and closed-source LLMs with up to 15.3% relative improvement in task utility and 9.3% in privacy–utility trade-off over baselines.

2 Related Work

Multi-agent collaboration organizes multiple LLM agents to exchange information and coordinate their capabilities toward shared objectives (Tran et al., 2025; Li et al., 2024; Dang et al., 2026). Depending on where such information exchange occurs in the pipeline, existing collaboration mechanisms can be broadly categorized into early-, mid-, and late-stage paradigms (Tran et al., 2025):

2.1 EARLY-STAGE COLLABORATION

Early-stage collaboration occurs before agent reasoning and mainly shares data, context, environmental observations, or input representations among agents (Wang et al., 2026a; Rezazadeh et al., 2025; Lu et al., 2025; Li et al., 2023). For example, Collaborative Memory maintains shared memory tiers, allowing selected memory fragments to be exposed across users and agents (Rezazadeh et al., 2025). Such direct memory sharing provides richer contextual information for collaboration, but also enlarges the exposure surface of underlying private data in cross-organizational settings.

2.2 MID-STAGE COLLABORATION

Mid-stage collaboration enables agents to interact through model-level information during training or inference, such as parameters, weights, or learned representations (Bonawitz et al., 2021; Ye et al., 2024; Saha et al., 2025). Federated Instruction Tuning (FedIT), for instance, collaboratively fine-tunes LLMs by aggregating model updates learned from decentralized instruction data (Zhang et al., 2024), while FLoRA further supports federated fine-tuning with heterogeneous LoRA adapters through model-side aggregation (Wang et al., 2024). Although these approaches avoid directly exchanging raw data, they typically require access to model parameters and compatible training interfaces, making them difficult to cross-organizational deployment and for joint training.

2.3 LATE-STAGE COLLABORATION

Late-stage collaboration mainly operates on task- or output-level information and is therefore more compatible with independently deployed LLM agents (Xu et al., 2025; Reid et al., 2026; Chen et al., 2026; Hui et al., 2026; Tang et al., 2024). For task sharing, AgentVerse coordinates multiple experts around a shared task goal (Chen et al., 2024), while PromptObfus obfuscates sensitive query content with prediction-preserving alternatives (Li et al., 2026). These approaches highlight a fundamental tension in task-level collaboration: sharing richer task context increases information exposure, whereas restricting the shared context may weaken reasoning.

For output sharing, PrivCollab-Agent exchanges privacy-filtered claim-level messages instead of raw private contexts (Jiang et al., 2026), while RAGentA communicates document-grounded answers and explicit citations among specialized agents (Besrour et al., 2025). Therefore, final answers enable concise communication but provide limited collaborative context, while explicit evidence offers interpretable support but may not fully capture latent expertise.

3 PROBLEM FORMULATION

System and Reciprocal Communication. We consider a Multi-agent collaboration of independent LLM agents $\mathcal{A} = \{A_1, \ldots, A_N\}$, where each agent A_i maintains a private memory $\mathcal{M}_i = \{m_i^1, \ldots, m_i^{|\mathcal{M}_i|}\}$ (Tran et al., 2025; Dang et al., 2026). Given a query q with candidate answers $\mathcal{Y}$, one agent A_t serves as the *target agent*, while the remaining A_p act as *peer agents*. The target sends task information $q_{t\to p}$, either as the full query or a fragmented version, and each peer returns $A_p(q_{t\to p}, \mathcal{M}_p)$, typically as a final answer or explicit evidence. The target then predicts

$$\hat{y} = A_t(q, \mathcal{M}_t, A_p(q_{t\to p}, \mathcal{M}_p)), \qquad \hat{y} \in \mathcal{Y}. \tag{1}$$

Reconstruction Threat Model. We consider reconstruction attacks (Fredrikson et al., 2014; Staab et al., 2024) in which an agent attempts to infer private information from messages observed during normal collaboration. Given an observed message x, a reconstruction adversary produces

$$\hat{\mathcal{F}} = \mathrm{Fact}(x), \tag{2}$$

where $\hat{\mathcal{F}}$ denotes the private facts inferred about the sender. Privacy is therefore characterized by how much of the sender's underlying private information can be reconstructed from the exchanged message. Detailed attack and evaluation protocols are provided in Sec. C.

Heterogeneous Private Knowledge. Independent agents acquire private knowledge from different sources (Shi et al., 2025; Saha et al., 2025; Jin et al., 2021), leading to uneven coverage across domains. For example, one medical institution may hold richer cardiovascular cases, while another

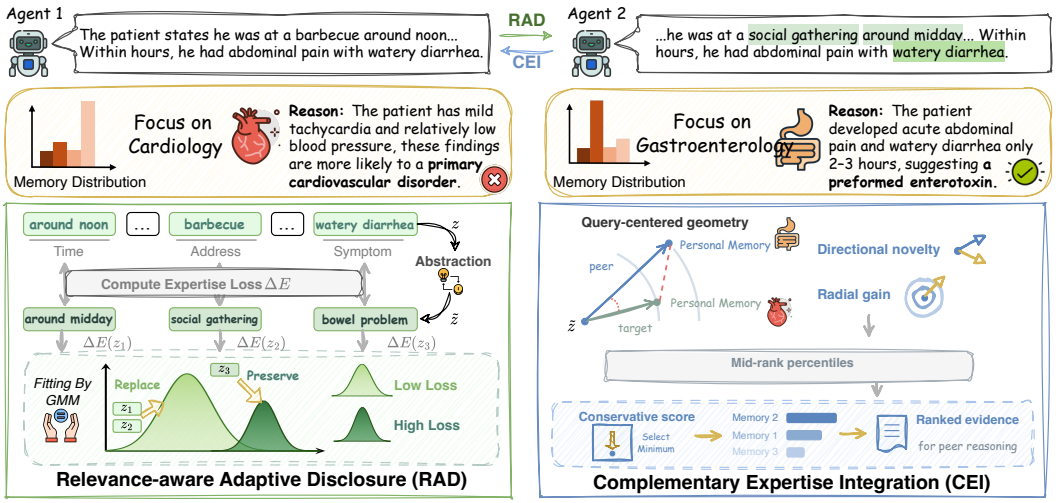


Figure 3: Overview of our proposed framework. Agent 1 applies RAD to obtain a privacy-controlled query and sends it to Agent 2. Agent 2 then uses CEI to identify complementary private knowledge and returns a compact reasoning summary to Agent 1. RAD enables adaptive disclosure without a fixed global threshold, while CEI captures complementarity through query-centered geometry.

has stronger gastrointestinal coverage. Let $\mathcal{C}$ denote the set of knowledge domains and $P_i(c)$ the proportion of agent A_i's private memory belonging to domain $c \in \mathcal{C}$. We refer to the resulting differences $P_i \neq P_j$ across agents as ***heterogeneous private knowledge***.

Heterogeneity-Guided Reciprocal Communication. The two communication boundaries require different decisions under heterogeneous private knowledge. When sending task information to peers, the target agent must limit sensitive disclosure while preserving valuable information. Let $\mathcal{Z}(q) = \{z_1, \ldots, z_L\}$ denote the sensitive spans in q, and $d_l \in \{0, 1\}$ the disclosure decision, where $d_l = 1$ preserves z_l and $d_l = 0$ replaces it with an abstraction. We formulate this trade-off as

$$\min_{\mathbf{d}} \left(\underbrace{\frac{1}{L}\sum_{l=1}^{L} d_l}_{\text{Privacy Exposure}} , \underbrace{\frac{1}{L}\sum_{l=1}^{L}(1-d_l)\Delta E_t(z_l)}_{\text{Expertise Loss}} \right), \tag{3}$$

where $\Delta E_t(z_l)$ denotes the expertise loss caused by abstracting z_l with respect to $\mathcal{M}_t$. This trade-off motivates that *Task disclosure should account for the value of information under heterogeneous private knowledge.* When peer knowledge returns, relevance alone is insufficient because it may still overlap with what the target already knows. We therefore characterize returned knowledge by what it contributes beyond the target's existing knowledge. This motivates that *Returned knowledge should be selected for its complementary value rather than absorbed indiscriminately.*

4 Methodology

4.1 Relevance-aware Adaptive Disclosure (RAD)

RAD balances sensitive query protection with preserving information relevant to the target agent's private expertise. Rather than uniformly masking sensitive information, it evaluates the expertise relevance lost when candidate abstractions are applied and makes query-local disclosure decisions accordingly. More details are shown in algorithm 2 in Sec. B.

Sensitive Span Extraction. Given a raw query q, we identify a set of sensitive spans

$$\mathcal{Z}(q) = \{z_1, \ldots, z_L\}. \tag{4}$$

These include universal personally identifiable information, such as `names`, `ages`, `locations`, `organizations`, and `identifiers`, as well as domain-specific sensitive attributes in medicine and law (Wang et al., 2026b; Pilán et al., 2022; Stubbs et al., 2015).

Candidate Semantic Abstraction. For each sensitive span $z \in \mathcal{Z}(q)$, we construct a candidate abstraction $\tilde{z}$ that removes unnecessary specificity while retaining task-relevant semantics. Whether $\tilde{z}$ replaces the original span is determined by the expertise relevance lost through abstraction.

Agent-Local Expertise Loss. We measure the relevance of z and $\tilde{z}$ to the target agent's private memory as

$$E_t(x) = \max\left(0, \max_{m_t^j \in \mathcal{M}_t} \mathrm{sim}\Big(f(x), f(m_t^j)\Big)\right), \quad x \in \{z, \tilde{z}\}. \tag{5}$$

where $E_t(\cdot)$ serves as a proxy for the association between the span and knowledge stored in the target agent's private memory.

The abstraction-induced expertise loss is then defined as

$$\Delta E_t(z) = \frac{E_t(z) - E_t(\tilde{z})}{E_t(z) + E_t(\tilde{z}) + \epsilon}, \tag{6}$$

where ϵ ensures numerical stability. Larger $\Delta E_t(z)$ indicates greater expertise relevance loss. The target memory serves as a domain-grounded reference for abstraction, allowing sensitive details to be generalized while preserving task-relevant domain semantics. This keeps the protected query informative for collaboration while making the original private information harder to reconstruct.

Query-Local Adaptive Disclosure. Since expertise-loss distributions vary across queries and agents, a fixed global threshold cannot reliably characterize what information is worth disclosing. We instead fit a two-component one-dimensional Gaussian mixture model (GMM) to the query-local expertise losses $\{\Delta E_t(z)\}_{z \in \mathcal{Z}(q)}$:

$$p(\Delta E) = \sum_{c \in \{\mathrm{low}, \mathrm{high}\}} \pi_c \mathcal{N}\big(\Delta E; \mu_c, \sigma_c^2\big), \qquad \mu_{\mathrm{low}} < \mu_{\mathrm{high}}. \tag{7}$$

The mixture induces a self-calibrated disclosure boundary through posterior inference:

$$\mathrm{Action}(z) = \begin{cases} \texttt{preserve}, & P(\mathrm{high} \mid \Delta E_t(z)) > P(\mathrm{low} \mid \Delta E_t(z)), \\ \texttt{abstract}, & \text{otherwise.} \end{cases} \tag{8}$$

This posterior rule is Bayes-optimal under the fitted mixture, yielding a query-specific boundary without manually specified thresholds; we provide the formal analysis in Sec. E.3. When the mixture cannot be reliably estimated, RAD adopts a privacy-first fallback and abstracts all sensitive spans.

4.2 Complementary Expertise Identification (CEI)

While RAD controls what information is disclosed, CEI identifies which peer knowledge is complementary to the target. Since retrieval relevance may still reflect redundant information, we measure complementarity relative to the target agent's knowledge. More details are shown in algorithm 3.

Query-Relative Memory Geometry. Let

$$\hat{\mathbf{q}} = \frac{f(\tilde{q})}{\|f(\tilde{q})\|}, \qquad \hat{\mathbf{m}} = \frac{f(m)}{\|f(m)\|}. \tag{9}$$

For each retrieved memory m, we characterize its position relative to the query by

$$\begin{aligned} \theta(m) &= \arccos\left(\hat{\mathbf{q}}^\top \hat{\mathbf{m}}\right), \\ \mathbf{u}(m) &= \frac{\hat{\mathbf{m}} - (\hat{\mathbf{q}}^\top \hat{\mathbf{m}})\hat{\mathbf{q}}}{\|\hat{\mathbf{m}} - (\hat{\mathbf{q}}^\top \hat{\mathbf{m}})\hat{\mathbf{q}}\|}. \end{aligned} \tag{10}$$

Here, $\theta(m)$ denotes the angular radius from the query, while $\mathbf{u}(m)$ captures the tangent semantic direction along which the memory extends beyond the query representation.

Pairwise Complementarity. Consider a peer memory m_p^j and a target memory m_t^i. We quantify their directional novelty and radial gain as

$$n_{ji} = \frac{\arccos\left(\mathbf{u}(m_p^j)^\top \mathbf{u}(m_t^i)\right)}{\pi}, \qquad g_{ji} = \max\left(0,\ \frac{\theta(m_p^j) - \theta(m_t^i)}{\theta(m_p^j)}\right). \tag{11}$$

The former measures whether the peer contributes a distinct semantic direction, while the latter captures whether it extends farther from the query than the corresponding target evidence.

To account for their different empirical distributions, we transform n_{ji} and g_{ji} into their respective mid-rank percentiles, $\bar{n}_{ji}$ and $\bar{g}_{ji}$, over all peer–target memory pairs. Their joint complementarity is then defined as

$$s_{ji} = 1 - (1 - \bar{n}_{ji})(1 - \bar{g}_{ji}). \tag{12}$$

Personalized Peer Evidence. For each peer memory m_p^j, we aggregate its pairwise complementarity by $s_p = \min_i s_{ji}$. The minimum aggregation imposes a conservative criterion: a peer memory receives a high importance score only when it remains complementary even with respect to the target memory that best covers it. Peer memories are subsequently ranked according to s_p and provided to the peer agent as personalized evidence for reasoning.

4.3 Privacy-Controlled Collaborative Reasoning

After RAD produces the privacy-controlled query $\tilde{q}$, the target retains the original query q for local reasoning and sends only $\tilde{q}$ to peer agents. Each agent retrieves its top-K private memories by

$$\mathcal{R}_i(q_i) = \mathrm{TopK}_{m \in \mathcal{M}_i} \mathrm{sim}(f(q_i), f(m)), \qquad q_i = \begin{cases} q, & i = t, \\ \tilde{q}, & i \neq t, \end{cases} \tag{13}$$

where $f(\cdot)$ denotes the embedding model and $\mathrm{sim}(\cdot,\cdot)$ is cosine similarity.

For each peer A_p, CEI evaluates the retrieved knowledge by its complementarity to the target agent and guides the generation of a compact, answer-free reasoning summary:

$$r_p = A_p\left(\tilde{q}, \left\{\left(m_p^j, s_p(m_p^j)\right)\right\}_{j=1}^{K}\right), \tag{14}$$

where $m_p^j \in \mathcal{R}_p(\tilde{q})$ and $s_p(m_p^j)$ denotes its complementarity score. The retrieved memories remain local and are not directly transmitted.

The target agent finally integrates the returned peer summaries with its own retrieved evidence:

$$\hat{y} = A_t\left(q, \mathcal{R}_t(q), \{r_p\}_{p \neq t}\right), \qquad \hat{y} \in \mathcal{Y}. \tag{15}$$

The complete workflow is illustrated in Fig. 3 and summarized in algorithm 1 of Sec. B, showing how RAD controls what leaves the target agent and CEI determines what complementary expertise is carried back for final reasoning.

5 Experiment

Dataset Construction.

We construct evaluation datasets for two privacy-sensitive domains, with category-level organization to support heterogeneous private-memory construction.

- **MedQA.** We construct a six-category subset from the English MedQA–USMLE dataset (Jin et al., 2021), with the evaluation set primarily based on MedQA-Hard (Tang et al., 2025). The training split is used to construct agents' private memories, while the test split is reserved for evaluation.
- **LegalQA.** We reformulate LegalQA (Yang et al., 2020) into a multiple-choice format and consolidate its topics into six legal categories. Likewise, the training split is used for private-memory construction and the test split for evaluation.

Table 1: Comparison on MedQA and LegalQA. The last column reports average TO. MP $= 100$ indicates no peer communication, whereas QP $= 0$ denotes full exposure of the query. Accordingly, **bold** values indicate the best results within the *Privacy Protecting + Collaboration* setting.

Methods	MedQA					LegalQA					Avg.
	TU(↑)	QP(↑)	MP(↑)	OP(↑)	TO(↑)	TU(↑)	QP(↑)	MP(↑)	OP(↑)	TO(↑)	TO(↑)
Qwen3-30B											
Single Agent	81.88	0.00	100.00	94.86	87.89	42.86	0.00	100.00	92.80	58.64	73.26
Privacy Protecting (without peer)											
MNI-Gate	69.90	63.75	100.00	97.27	81.35	38.10	20.79	100.00	96.03	54.55	67.95
PromptObfus	78.96	45.87	100.00	96.12	86.70	41.60	19.13	100.00	95.60	57.98	72.34
RAD	82.85	50.21	100.00	96.51	89.16	42.61	23.78	100.00	96.23	59.06	74.11
Collaboration (with peer)											
Reasoning Share	82.20	0.00	97.64	92.59	87.09	43.61	0.00	97.08	90.02	58.76	72.92
PrivCollab-Agent	82.52	0.00	98.31	93.23	87.55	39.85	0.00	97.60	90.50	55.33	71.44
CEI	83.50	0.00	96.25	91.27	87.21	44.61	0.00	94.74	87.81	59.17	73.19
Privacy Protecting + Collaboration											
MNI-Gate + Reasoning Share	81.23	62.11	97.71	94.91	87.54	39.35	28.60	97.09	93.70	55.42	71.48
PromptObfus + Reasoning Share	81.23	45.87	97.35	93.57	86.96	39.85	19.13	97.31	93.00	55.79	71.38
RAD + PrivCollab-Agent	82.52	50.21	98.80	95.38	88.49	42.86	23.78	98.64	94.92	59.05	73.77
Ours	84.79	50.21	96.37	93.03	**88.72**	45.36	23.78	94.69	91.23	**60.60**	**74.66**
Methods	**MedQA**					**LegalQA**					**Avg.**
	TU(↑)	QP(↑)	MP(↑)	OP(↑)	TO(↑)	TU(↑)	QP(↑)	MP(↑)	OP(↑)	TO(↑)	TO(↑)
DeepSeek-v4-Flash											
Single Agent	95.47	0.00	100.00	93.07	94.25	50.38	0.00	100.00	91.37	64.94	79.60
Privacy Protecting (without peer)											
MNI-Gate	89.97	58.51	100.00	95.85	92.82	47.12	21.46	100.00	95.34	63.07	77.94
PromptObfus	92.56	46.17	100.00	96.14	94.32	47.37	21.36	100.00	95.93	63.42	78.87
RAD	94.50	48.11	100.00	95.09	94.79	49.62	21.01	100.00	94.92	65.17	79.98
Collaboration (with peer)											
Reasoning Share	95.79	0.00	97.21	90.45	93.05	50.38	0.00	97.53	89.24	64.40	78.72
PrivCollab-Agent	95.79	0.00	99.05	92.18	93.95	49.12	0.00	97.93	89.64	63.47	78.71
CEI	97.09	0.00	97.20	90.51	93.68	50.75	0.00	95.67	87.55	64.26	78.97
Privacy Protecting + Collaboration											
MNI-Gate + Reasoning Share	92.23	57.98	97.46	93.23	92.73	48.62	21.46	97.30	93.08	63.88	78.30
PromptObfus + Reasoning Share	94.17	46.17	97.57	93.82	94.00	48.12	21.36	97.85	93.89	63.63	78.81
RAD + PrivCollab-Agent	94.82	48.11	98.89	94.04	**94.43**	48.62	21.01	97.77	93.07	63.87	79.15
Ours	96.44	48.11	96.65	91.96	94.15	49.62	20.92	95.57	91.03	**64.23**	**79.19**

LLM Backbones. Given the privacy-sensitive setting, we primarily consider open-source LLMs, including DeepSeek-V4-Flash (Xu et al., 2026) and Qwen3-30B (Yang et al., 2025), which can support deployment under more controlled inference environments. We additionally evaluate GPT-4o-mini (Hurst et al., 2024) as a closed-source LLM to examine whether the effectiveness of our framework generalizes across different model access settings.

Memory Construction. For each dataset, we partition the training set across three clients using a Dirichlet distribution with $\alpha = 0.5$, following common non-IID data partitioning protocols in federated learning (Ng et al., 2011; Zhang et al., 2022; 2024). The resulting category-skewed private memories induce heterogeneous expertise across agents.

Evaluation Metrics. We use five metrics: *Task Utility (TU)*, *Query Protection (QP)*, *Memory Protection (MP)*, *Overall Protection (OP)*, and *Trade-off (TO)* (Chu et al., 2024; Park et al., 2026; Min et al., 2023; Honovich et al., 2022). TU measures answer accuracy; QP and MP measure protection of query and private-memory facts; OP captures their joint protection; and TO summarizes the privacy–utility trade-off. See Sec. C for detailed definitions.

Counterparts. We use *Single Agent* as the non-collaborative baseline (Yao et al., 2023). For privacy protection without peer collaboration, we compare with *MNI-Gate* (Chen et al., 2026) , *PromptObfus* (Li et al., 2026), and our RAD. For collaboration without query protection, we include *Reasoning Share* (Tang et al., 2024), *PrivCollab-Agent* (Jiang et al., 2026), and our CEI. We further evaluate their privacy–collaboration combinations, while **Ours** integrates RAD and CEI.

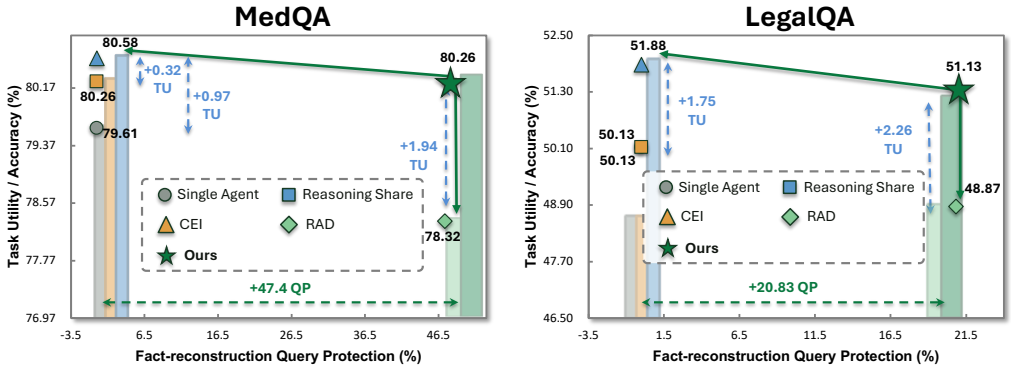
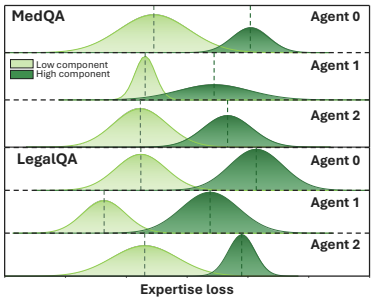


Figure 4: Privacy–Utility trade-off of different collaboration strategies on MedQA and LegalQA with GPT-4o-mini. CEI improves task utility, while RAD substantially increases query protection; combining them retains the protection gain while recovering task utility.



(a) GMM distributions.

Dataset	Criterion	TU(↑)	QP(↑)	TO(↑)
MedQA	Zero Threshold	80.58	**53.56**	87.92
	Mixture Mean	81.23	51.94	88.25
	Median	81.88	52.72	88.65
	GMM Posterior	**82.85**	50.21	**89.16**
LegalQA	Zero Threshold	40.20	**26.27**	56.76
	Mixture Mean	41.84	23.02	58.31
	Median	41.58	23.60	58.08
	GMM Posterior	**42.61**	23.78	**59.06**

(b) Comparison of alternative disclosure criteria.

Figure 5: Analysis of GMM modeling on MedQA and LegalQA. (a) Fitted low and highloss GMM components vary across agents and datasets. (b) GMM-based disclosure achieves the strongest task utility and overall privacy–utility trade-off compared with alternative threshold criteria.

5.1 Comparison across Privacy and Collaboration Settings

Overall Comparison. As shown in Table 1, the two modules exhibit complementary effects. RAD substantially improves query protection while preserving strong task utility, whereas CEI consistently improves task utility by exploiting complementary peer expertise, at the cost of some memory exposure. Combining both modules yields the strongest collaborative results. For Qwen3-30B, Ours achieves the highest TU on both MedQA and LegalQA and improves the average TO to 74.66, outperforming the two alternative privacy–collaboration combinations. For DeepSeek-V4-Flash, Ours likewise achieves the highest average TO within the *Privacy Protecting + Collaboration* setting (79.19), while substantially improving TU over the privacy-only alternatives. These results show that adaptive query disclosure and complementary expertise utilization jointly provide a more favorable privacy–utility balance across datasets and model backbones.

Effectiveness on Closed-Source LLMs. As shown in Fig. 4, CEI consistently improves task utility over standard collaboration, yielding gains of $+0.32$ on MedQA and $+1.75$ on LegalQA. Applying RAD substantially improves query protection, but incurs some utility degradation. Our full framework combines both advantages: compared with RAD alone, it recovers $+1.94$ TU on MedQA and $+2.26$ TU on LegalQA while retaining the improved query protection. These results indicate that the proposed framework remains effective with the closed-source GPT-4o-mini backbone.

5.2 Diagnostic Analysis

Analysis of GMM-based Disclosure. As shown in Fig. 5, the fitted low- and high-loss components are clearly distinguishable across agents and datasets, motivating query-local separation rather than

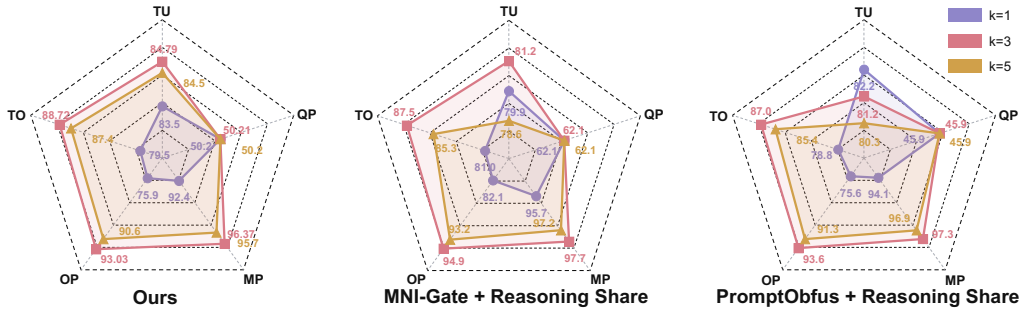


Figure 6: Sensitivity to retrieval depth K on Ours and two methods. We compare $K \in \{1, 3, 5\}$ over all metrics; $K = 3$ provides the most balanced privacy–utility performance across methods.

a fixed global threshold. Quantitatively, the GMM posterior achieves the highest TU and TO on both datasets. Although the zero threshold yields higher QP, it sacrifices more task utility, while the mean and median criteria provide weaker overall trade-offs. These results indicate that posterior-based separation better balances expertise preservation and query protection than simple thresholding rules. More results can be seen in Sec. E and Fig. 9.

Effect of Retrieval Depth. We vary the retrieval depth $K \in \{1, 3, 5\}$ and observe that $K = 3$ consistently provides the strongest overall privacy–utility balance across different collaboration strategies. For Ours, $K = 3$ achieves the highest TO (87.9), while $K = 5$ only slightly improves TU but reduces OP and TO. Similar trends hold for the two combinations. We therefore set $K = 3$ throughout our experiments.

Effect of the Number of Agents. We vary the number of agents from 3 to 5 and observe performance across different collaboration scales, with TU ranging only from 83.25 to 83.50. In contrast, the two methods are more sensitive to additional agents, particularly MNI-Gate + Reasoning Share, whose TU decreases from 81.23 to 76.89. Since three agents already provide consistent performance across methods, we adopt 3 agents in all experiments. More results can be seen in Sec. F.

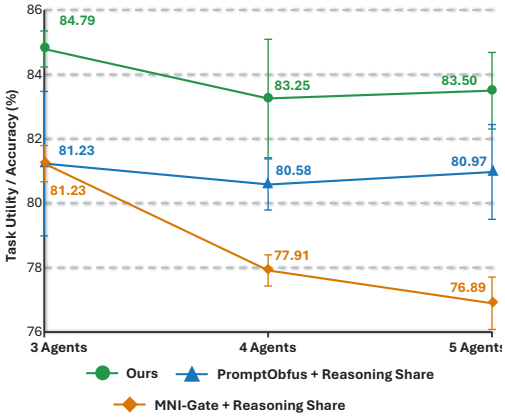


Figure 7: Number of agents' effect on task utility. Ours remains stable as the number of agents increases, while the baseline combinations exhibit larger performance degradation.

Effect of Private-Memory Heterogeneity. We further examine how heterogeneous private memories affect agent expertise and collaboration. Across GPT-4o-mini and Qwen3-30B, different memory coverage induces distinct agent strengths and changes single-agent performance across heterogeneity levels. Moreover, direct reasoning sharing provides only limited gains, whereas CEI consistently benefits more from complementary peer expertise. These results support our motivation that effective collaboration should account for both uneven local knowledge coverage and the complementarity among agents. Detailed qualitative and quantitative evidence is provided in Sec. A.

6 Conclusion

We suggest a broader perspective for privacy-aware Multi-agent collaboration: knowledge heterogeneity across institutions can serve not only as a source of complementary expertise, but also as a basis for controlling how that expertise is exchanged. In healthcare, this can support cross-specialty consultation while limiting unnecessary exposure of patient information; in legal settings, it can enable expertise sharing across firms while preserving sensitive client and case information. More broadly, privacy-aware collaboration need not rely on uniformly restricting communication. Instead, communication can be shaped by what each institution knows and what additional expertise others can contribute, opening a path toward more practical cross-organizational agent collaboration.

Reproducibility Statement

We provide detailed descriptions and pseudocode for the overall framework, RAD, and CEI in the main paper and appendix. The appendix further documents dataset construction, evaluation metrics, the query-local GMM formulation and proof, as well as ablations on key design choices. Model backbones, retrieval settings, and experimental configurations are specified to facilitate reproduction of all reported results.

AI Use Statement

Generative AI tools were used to assist with writing and language polishing, research execution, and the derivation and checking of mathematical arguments. The authors independently verified all references, mathematical claims, experimental results, and research conclusions, and take full responsibility for the content of the manuscript.

References

Ines Besrour, Jingbo He, Tobias Schreieder, and Michael Färber. Ragenta: Multi-agent retrieval-augmented generation for attributed question answering. *arXiv preprint arXiv:2506.16988*, 2025.

Kallista Bonawitz, Peter Kairouz, Brendan McMahan, and Daniel Ramage. Federated learning and privacy: Building privacy-preserving systems for machine learning and data science on decentralized data. *ACM Queue*, 2021.

Jinyu Chen, Feiyang Wang, Tian Guan, Yumeng Ma, Linghao Yang, and Yutong Wang. Min-trust: A minimum necessary information trust orchestration framework for multi-agent collaboration. In *Proceedings of the International Conference on Generative Artificial Intelligence and Education*, pp. 16–23, 2026.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, et al. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *International Conference on Learning Representations*, 2024.

Junjie Chu, Zeyang Sha, Michael Backes, and Yang Zhang. Reconstruct your previous conversations! comprehensively investigating privacy leakage risks in conversations with gpt models. pp. 6584–6600, 2024.

Yufan Dang, Chen Qian, Xueheng Luo, Jingru Fan, Zihao Xie, Ruijie Shi, Weize Chen, Cheng Yang, Xiaoyin Che, Ye Tian, et al. Multi-agent collaboration via evolving orchestration. *Conference on Neural Information Processing Systems*, 2026.

Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: An {End-to-End} case study of personalized warfarin dosing. In *USENIX Security Symposium*, pp. 17–32, 2014.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. True: Re-evaluating factual consistency evaluation. pp. 3905–3920, 2022.

Zheng Hui, Yijiang River Dong, Sanhanat Sivapiromrat, Ehsan Shareghi, and Nigel Collier. Privacy-r1: Privacy-aware multi-llm agent collaboration via reinforcement learning. 2026.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Yu Jiang, Zhengxi Xiao, Ourong Lin, Zifeng Gu, and Zhiqing Yang. Privcollab-agent: Privacy-preserving multi-agent collaboration under partial information. 2026.

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. *Applied Sciences*, 11(14):6421, 2021.

Huao Li, Yu Chong, Simon Stepputtis, Joseph P Campbell, Dana Hughes, Charles Lewis, and Katia Sycara. Theory of mind for multi-agent collaboration via large language models. In *Conference on Empirical Methods in Natural Language Processing*, 2023.

Xinyi Li, Sai Wang, Siqi Zeng, Yu Wu, and Yi Yang. A survey on llm-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth*, 1(1):9, 2024.

Xuan Li, Zhe Yin, Xiaodong Gu, and Beijun Shen. Anti-adversarial learning: Desensitizing prompts for large language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 31787–31795, 2026.

Jie Liu, Wenxuan Wang, Zizhan Ma, Guolin Huang, Yihang Su, Kao-Jung Chang, Wenting Chen, Haoliang Li, Linlin Shen, and Michael Lyu. Medchain: Bridging the gap between llm agents and clinical practice with interactive sequence. 2025.

Kai Lu, Shixiong Xu, Jinqiu Li, Kun Ding, and Gaofeng Meng. Agent reviewers: Domain-specific multimodal agents with shared memory for paper review. In *International Conference on Machine Learning*, volume 267, pp. 40803–40830. PMLR, 2025.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, 2023.

Kai Wang Ng, Guo-Liang Tian, and Man-Lai Tang. Dirichlet and related distributions: Theory, methods and applications. 2011.

Minjun Park, Donghyun Kim, Hyeonjong Ju, Seungwon Lim, Dongwook Choi, Taeyoon Kwon, Minju Kim, and Jinyoung Yeo. Pac-bench: Evaluating multi-agent collaboration under privacy constraints. pp. 31030–31056, 2026.

Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. The text anonymization benchmark (tab): A dedicated corpus and evaluation framework for text anonymization. *Computational Linguistics*, 48(4):1053–1101, 2022.

Alistair Reid, Simon O'Callaghan, Dustin Venini, Liam Carroll, and Tiberio Caetano. Risks and controls for multi-agent systems: an analytical framework for deployment of ai agents across organisational boundaries. *arXiv preprint arXiv:2608.26626*, 2026.

Alireza Rezazadeh, Zichao Li, Ange Lou, Yuying Zhao, Wei Wei, and Yujia Bao. Collaborative memory: Multi-user memory sharing in llm agents with dynamic access control. *arXiv preprint arXiv:2505.18279*, 2025.

Pramit Saha, Divyanshu Mishra, Felix Wagner, Konstantinos Kamnitsas, and J Alison Noble. Incongruent multimodal federated learning for medical vision and language-based multi-label disease detection. In *AAAI Conference on Artificial Intelligence*, volume 39, pp. 28331–28339, 2025.

Zitong Shi, Guancheng Wan, Wenke Huang, Guibin Zhang, Jiawei Shao, Mang Ye, and Carl Yang. Privacy-enhancing paradigms within federated multi-agent systems. *arXiv preprint arXiv:2503.08175*, 2025.

Robin Staab, Mark Vero, Mislav Balunovic, and Martin Vechev. Beyond memorization: Violating privacy via inference with large language models. In *International Conference on Learning Representations*, volume 2024, pp. 33832–33878, 2024.

Amber Stubbs, Christopher Kotfila, and Özlem Uzuner. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/uthealth shared task track 1. *Journal of biomedical informatics*, 58:S11–S19, 2015.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. Medagents: Large language models as collaborators for zero-shot medical reasoning. 2024. Findings.

Xiangru Tang, Daniel Shao, Jiwoong Sohn, Jiapeng Chen, Jiayi Zhang, Jinyu Xiang, Fang Wu, Yilun Zhao, Chenglin Wu, Wenqi Shi, et al. Medagentsbench: Benchmarking thinking models and agent frameworks for complex medical reasoning. *arXiv e-prints*, pp. arXiv–2503, 2025.

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D Nguyen. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322*, 2025.

Dong Wang, Yang Li, Ansong Ni, Ching-Feng Yeh, Youssef Emad, Xinjie Lei, Liam Robbins, Karthik Padthe, Hu Xu, Xian Li, et al. Matrix: Peer-to-peer multi-agent synthetic data generation framework. In *Conference on Machine Learning and Systems*, 2026a.

Lei Wang, Min Huang, and Eduard Dragut. Danceha: A multi-agent framework for document-level aspect-based sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 29714–29722, 2026b.

Wenxuan Wang, Zizhan Ma, Zheng Wang, Chenghan Wu, Wenting Chen, Xiang Li, and Yixuan Yuan. A survey of llm-based agents in medicine: How far are we from baymax? 2025.

Ziyao Wang, Zheyu Shen, Yexiao He, Guoheng Sun, Hongyi Wang, Lingjuan Lyu, and Ang Li. Flora: Federated fine-tuning large language models with heterogeneous low-rank adaptations. *Conference on Neural Information Processing Systems*, 37:22513–22533, 2024.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Wenchao Xu, Jinyu Chen, Peirong Zheng, Xiaoquan Yi, Tianyi Tian, Wenhui Zhu, Quan Wan, Haozhao Wang, Yunfeng Fan, Qinliang Su, et al. Deploying foundation model powered agent services: A survey. *IEEE Communications Surveys & Tutorials*, 28:1483–1519, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Min Yang, Lei Chen, Ziyu Lyu, Junhao Liu, Ying Shen, and Qingyao Wu. Hierarchical fusion of common sense knowledge and classifier decisions for answer selection in community question answering. *Neural Networks*, 132:53–65, 2020.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

Rui Ye, Wenhao Wang, Jingyi Chai, Dihan Li, Zexi Li, Yinda Xu, Yaxin Du, Yanfeng Wang, and Siheng Chen. Openfedllm: Training large language models on decentralized private data via federated learning. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024.

Shengbin Yue, Ting Huang, Zheng Jia, Siyuan Wang, Shujun Liu, Yun Song, Xuan-Jing Huang, and Zhongyu Wei. Multi-agent simulator drives language models for legal intensive interaction. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 6552–6585, 2025.

Jianyi Zhang, Saeed Vahidian, Martin Kuo, Chunyuan Li, Ruiyi Zhang, Tong Yu, Guoyin Wang, and Yiran Chen. Towards building the federated gpt: Federated instruction tuning. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024.

Zhuo Zhang, Yuanhang Yang, Yong Dai, Lizhen Qu, and Zenglin Xu. When federated learning meets pre-trained language models' parameter-efficient tuning methods. *arXiv preprint arXiv:2212.10025*, 2022.

APPENDIX

A Evidence for Heterogeneous Expertise. **15**

B Algorithmic Overview. **15**

C Evaluation Metrics **16**

D Details of Comparison Methods **20**

E Analysis of Query-wise GMM Calibration **22**

E.1 Compared Disclosure Criteria 22

E.2 Posterior Decision Rule 22

E.3 Bayes Optimality 23

E.4 Existence of an Interior Calibration Boundary 23

E.5 Closed-form Boundary under Equal Variance 24

E.6 Query-wise Self-Calibration 24

E.7 Consistency of the Estimated Boundary 25

E.8 Qualitative Analysis of Adaptive Disclosure 26

F Effect of the Number of Agents. **26**

Table 2: Summary of key notations.

Category	Notation	Description
General	$\mathcal{A} = \{A_1, \ldots, A_N\}$	Set of N independently operated LLM agents.
	A_i	The i-th agent in the multi-agent system.
	$A_t,\ A_p$	Target agent and a peer agent, respectively.
	$\mathcal{M}_i$	Private memory associated with agent A_i.
	m_i^j	The j-th memory item in $\mathcal{M}_i$.
	$q,\ \tilde{q}$	Original query and its privacy-controlled representation.
	$\mathcal{Y}$	Set of candidate answers for query q.
	$\mathcal{R}_i(x)$	Top-K memories retrieved by agent A_i using query x.
	$f(\cdot)$	Embedding function.
	$\mathrm{sim}(\cdot,\cdot)$	Cosine similarity between two embeddings.
RAD	$\mathcal{Z}(q) = \{z_1, \ldots, z_L\}$	Set of sensitive spans detected in query q.
	$z,\ \tilde{z}$	Sensitive span and its candidate semantic abstraction.
	$d_l,\ \mathbf{d}$	Disclosure decision for z_l and the corresponding decision vector.
	$E_t(x)$	Expertise relevance of x to the target agent's private memory.
	$\Delta E_t(z)$	Expertise relevance loss induced by abstracting z.
	π_c, μ_c, σ_c^2	Mixing weight, mean, and variance of GMM component c.
CEI	$\hat{\mathbf{q}},\ \hat{\mathbf{m}}$	Normalized embeddings of the protected query and a retrieved memory.
	$\theta(m)$	Angular radius of memory m relative to the protected query.
	$\mathbf{u}(m)$	Tangent direction of memory m relative to the protected query.
	$n_{ji},\ g_{ji}$	Directional novelty and radial gain between peer memory m_p^j and target memory m_t^i.
	$\bar{n}_{ji},\ \bar{g}_{ji}$	Mid-rank percentiles of n_{ji} and g_{ji}, respectively.
	s_{ji}	Pairwise complementarity between peer memory m_p^j and target memory m_t^i.
	s_p	Aggregated importance score associated with peer agent A_p.
	r_p	Compact reasoning summary generated by peer agent A_p.
	$\hat{y}$	Final prediction produced by the target agent.

A Evidence for Heterogeneous Expertise.

As shown in Fig. 8(a), the phenomenon observed in the main-text motivation is also present with GPT-4o-mini: agents with different private-memory coverage exhibit distinct category-level strengths, indicating that local knowledge composition directly shapes agent expertise. Fig. 8(b) further provides quantitative evidence with Qwen3-30B under different heterogeneity levels. The single-agent accuracy changes from 81.23% at $\alpha = 0.1$ to 80.91% at $\alpha = 1.0$, showing that memory partitioning alone affects agent-level capability. Moreover, direct Reasoning Share provides limited improvement ($+0.32$ and $+0.00$ points), whereas CEI consistently yields larger gains ($+1.94$ and $+4.20$ points). Together, these results support the view that heterogeneous private memories induce distinct agent expertise, and that explicitly identifying complementary knowledge is more effective than directly exchanging peer reasoning.

B Algorithmic Overview.

We provide pseudocode for the complete collaboration workflow and its two core components. Algorithm 1 summarizes the end-to-end procedure, where the target agent first applies RAD to obtain the privacy-controlled query, peer agents then retrieve and score complementary evidence through CEI, and the target finally integrates the returned reasoning summaries with its local evidence. Algorithms 2 and 3 detail the query-adaptive disclosure and complementary expertise identification procedures, respectively.

Table 3: Representative examples from the constructed MedQA and LegalQA datasets. We present two examples from each dataset, covering different medical and legal categories.

Dataset	Category	Question	Options	Answer
MedQA	Cardiology	A 43-year-old woman is brought to the emergency department by her brother for severe chest pain. The patient recently lost her husband in a car accident and is still extremely shocked by the event. On physical examination, her blood pressure is 105/67 mm Hg, the heart rate is 96/min and regular, breathing rate is 23/min, and the pulse oximetry is 96%. An S3 heart sound and rales in the lower right and left lung lobes are heard. A 12-lead ECG shows no significant findings. Echocardiography shows an enlarged left ventricle and left atrium. Which of the following is the most likely diagnosis?	(A) Atrial fibrillation (B) Constrictive pericarditis (C) Takotsubo cardiomyopathy (D) Restrictive cardiomyopathy	C
	Psychiatry	A 16-year-old girl is brought to the physician by her father because of concerns about her behavior during the past 2 years. She does not have friends and spends most of the time reading by herself. Her father says that she comes up with excuses to avoid family dinners and other social events. She states that she likes reading and feels more comfortable on her own. On mental status examination, her thought process is organized and logical. Her affect is flat. Which of the following is the most likely diagnosis?	(A) Schizoid personality disorder (B) Antisocial personality disorder (C) Schizophreniform disorder (D) Autism spectrum disorder	A
LegalQA	Contract & Commercial Law	The insurance people misled the old people to buy Tae Con. How should we deal with it?	(A) Hello, it's been so long that the time limit for the guarantee has passed. (B) This is a different matter and cannot be confused. Please accept this reply if you are satisfied. (C) In case of alleged insurance fraud, you have the right to file a complaint with the Insurance Supervision Department or to file a direct complaint with the court for the return of the alleged premium. (D) Refunds are required within 24 per cent of principal and legally supported annual interest rates.	C
	Public & Procedural Law	I was sentenced to two years in prison, and the trial was pending until the date of its entry into force, and the date of execution is almost ready.	(A) Hello, the bank's card bill is still due from you, and the maintenance can be negotiated, and the court can not take the child's custody. (B) One of the conditions for out-of-hospital execution is that there are serious illnesses requiring medical care, but the certificate must be issued by a hospital designated by the provincial people's government and approved in accordance with the law. (C) If the windows are not added by you, the lessor should bear the corresponding costs of the modification. (D) You're not going to get it, but I suggest you pay it back.	B

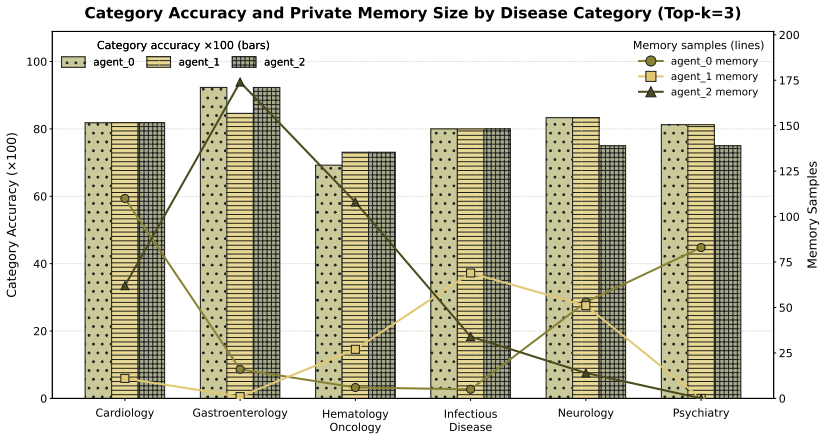
C Evaluation Metrics

We evaluate each method along two complementary dimensions: whether it solves the task correctly and how much information remains protected during inter-agent communication. All five metrics below lie in $[0, 1]$ and are reported as percentages; higher values are better.

Fact-level protection assessment. For each evaluation instance n, we extract atomic facts from the raw query and the retrieved peer memories (Min et al., 2023):

$$\mathcal{F}_n^Q = \mathrm{Fact}(q_n), \qquad \mathcal{F}_n^M = \mathrm{Fact}\left(\bigcup_{p \neq t} \mathcal{R}_p(\tilde{q}_n)\right). \tag{16}$$

Facts in $\mathcal{F}_n^M$ already entailed by q_n are excluded to avoid counting task information as private-memory leakage.



(a) Category-level performance and private-memory coverage across agents on MedQA with GPT-4o-mini.

α	Agent	Single Agent	Reasoning Share	CEI
0.1	Agent 0	83.50	84.47	**86.41**
	Agent 1	79.61	79.61	**80.58**
	Agent 2	80.58	80.58	**82.52**
	Overall	81.23	81.55 (+0.32)	**83.17** (+1.94)
1.0	Agent 0	78.64	79.61	**87.38**
	Agent 1	82.52	83.50	**85.44**
	Agent 2	81.55	79.61	**82.52**
	Overall	80.91	80.91 (+0.00)	**85.11** (+4.20)

(b) Qwen3-30B performance under different private-memory heterogeneity levels. Gains are measured relative to Single Agent.

Figure 8: Supplementary evidence of expertise induced by heterogeneous private memories. (a) With GPT-4o-mini, differences in local memory coverage correspond to distinct category-level strengths across agents. (b) With Qwen3-30B, varying private-memory heterogeneity changes both single-agent performance and the benefit of collaboration, while CEI consistently outperforms direct reasoning sharing.

For query protection, the attacker observes the peer-visible query representation, e.g., $(\tilde{q}_n, \mathcal{Y}_n)$ for our method. For memory protection, the attacker observes the target-visible context, $(q_n, \mathcal{Y}_n, \{r_{p,n}\}_{p \neq t})$.

Let $\widehat{\mathcal{F}}_n^Q$ and $\widehat{\mathcal{F}}_n^M$ denote the recovered query and memory facts, respectively. A source fact f is considered recovered if a reconstructed fact r entails it according to the fixed semantic judge, i.e., $J_{\text{ent}}(r, f) = 1$ (Honovich et al., 2022). Let N_{eval} denote the number of evaluated target–instance pairs.

Task Utility (TU). Task utility measures the accuracy of the final prediction $\hat{y}_n$:

$$\text{TU} = \frac{1}{N_{\text{eval}}} \sum_{n=1}^{N_{\text{eval}}} \mathbb{1}\left[\hat{y}_n = y_n^*\right], \tag{17}$$

where $y_n^* \in \mathcal{Y}_n$ denotes the ground-truth answer. For multi-answer questions, the predicted and ground-truth answer sets must match exactly (Park et al., 2026).

Algorithm 1: Privacy-Aware Multi-Agent Collaboration

Environment : Agent set $\mathcal{A} = \{A_1, \ldots, A_N\}$ with private memories $\{\mathcal{M}_i\}_{i=1}^{N}$; embedding function $f(\cdot)$.
Input : Raw query q with candidate answers $\mathcal{Y}$; target agent A_t; retrieval depth K.
Output : Final prediction $\hat{y}$.

// --- Target-Side Query Protection ---
$\tilde{q} \leftarrow RAD(q, \mathcal{M}_t)$
Retrieve target evidence using the original query:

$$\mathcal{R}_t(q) \leftarrow \mathrm{TopK}_{m \in \mathcal{M}_t} \mathrm{sim}\big(f(q), f(m)\big)$$

// --- Peer-Side Complementary Reasoning ---
foreach $A_p \in \mathcal{A} \setminus \{A_t\}$ **do**
Retrieve peer evidence using the protected query:

$$\mathcal{R}_p(\tilde{q}) \leftarrow \mathrm{TopK}_{m \in \mathcal{M}_p} \mathrm{sim}\big(f(\tilde{q}), f(m)\big)$$

Score complementary peer evidence:

$$\{(m_p^j, s_p)\}_{j=1}^{K} \leftarrow CEI\big(\tilde{q}, \mathcal{R}_t(q), \mathcal{R}_p(\tilde{q})\big)$$

Generate compact peer reasoning:

$$r_p \leftarrow A_p\left(\tilde{q}, \{(m_p^j, s_p)\}_{j=1}^{K}\right)$$

// --- Target-Side Final Reasoning ---
Combine local evidence and peer reasoning summaries:

$$\hat{y} \leftarrow A_t\left(q, \mathcal{R}_t(q), \{r_p\}_{p \neq t}\right)$$

return $\hat{y}$

Query Protection (QP). Query protection measures the fraction of raw-query facts that cannot be recovered from the peer-visible query representation:

$$\mathrm{QP}_n = 1 - \frac{\left|\widehat{\mathcal{F}}_n^Q\right|}{\left|\mathcal{F}_n^Q\right|}, \qquad \mathrm{QP} = \frac{1}{N_{\mathrm{eval}}} \sum_{n=1}^{N_{\mathrm{eval}}} \mathrm{QP}_n. \tag{18}$$

Memory Protection (MP). Memory protection measures the fraction of retrieved peer-memory facts that remain unrecoverable from the target-visible peer messages:

$$\mathrm{MP}_n = 1 - \frac{\left|\widehat{\mathcal{F}}_n^M\right|}{\left|\mathcal{F}_n^M\right|}, \qquad \mathrm{MP} = \frac{1}{N_{\mathrm{eval}}} \sum_{n=1}^{N_{\mathrm{eval}}} \mathrm{MP}_n. \tag{19}$$

Methods that transmit no peer-memory information recover no memory facts through this communication channel.

Overall Protection (OP). We jointly measure protection of query and peer-memory facts within each instance:

$$\mathrm{OP}_n = 1 - \frac{\left|\widehat{\mathcal{F}}_n^Q\right| + \left|\widehat{\mathcal{F}}_n^M\right|}{\left|\mathcal{F}_n^Q\right| + \left|\mathcal{F}_n^M\right|}, \tag{20}$$

$$\mathrm{OP} = \frac{1}{N_{\mathrm{eval}}} \sum_{n=1}^{N_{\mathrm{eval}}} \mathrm{OP}_n. \tag{21}$$

Thus, OP is fact-count weighted between query and memory protection within each instance, while each target–instance pair contributes equally to the dataset-level mean.

Algorithm 2: Relevance-aware Adaptive Disclosure (RAD)

Environment : Target private memory $\mathcal{M}_t$; embedding function $f(\cdot)$.
Input : Raw query q.
Output : Privacy-controlled query $\tilde{q}$.

// --- Sensitive Span Extraction ---
Detect sensitive spans:

$$\mathcal{Z}(q) = \{z_1, \ldots, z_L\}$$

// --- Agent-Local Expertise Loss ---
foreach $z \in \mathcal{Z}(q)$ **do**
  Generate candidate semantic abstraction $\tilde{z}$
  Compute expertise relevance:

$$E_t(x) \leftarrow \max\left(0,\ \max_{m \in \mathcal{M}_t} \mathrm{sim}\big(f(x), f(m)\big)\right), \quad x \in \{z, \tilde{z}\}$$

  Compute abstraction-induced expertise loss:

$$\Delta E_t(z) \leftarrow \frac{E_t(z) - E_t(\tilde{z})}{E_t(z) + E_t(\tilde{z}) + \epsilon}$$

// --- Query-Local Adaptive Disclosure ---
if $\{\Delta E_t(z)\}_{z \in \mathcal{Z}(q)}$ *supports a distinguishable two-component fit* **then**
  Fit the query-local GMM:

$$p(\Delta E) = \sum_{c \in \{\mathrm{low}, \mathrm{high}\}} \pi_c \mathcal{N}\left(\Delta E; \mu_c, \sigma_c^2\right)$$

  Order components such that

$$\mu_{\mathrm{low}} < \mu_{\mathrm{high}}$$

  foreach $z \in \mathcal{Z}(q)$ **do**
    if $P(\mathrm{high} \mid \Delta E_t(z)) > P(\mathrm{low} \mid \Delta E_t(z))$ **then**
      Preserve z
    else
      Replace z with $\tilde{z}$
else
  Abstract all detected sensitive spans // Conservative fallback
Construct $\tilde{q}$ from the resulting span decisions
return $\tilde{q}$

Privacy–Utility Trade-off (TO). We summarize the balance between task utility and overall protection using their harmonic mean:

$$\mathrm{TO} = \frac{2\,\mathrm{TU} \cdot \mathrm{OP}}{\mathrm{TU} + \mathrm{OP}}. \tag{22}$$

When $\mathrm{TU} = \mathrm{OP} = 0$, we define $\mathrm{TO} = 0$.

The harmonic mean is well suited to our setting because it:

- **treats utility and privacy symmetrically**, without favoring either objective;
- **penalizes imbalance**, as the score is dominated by the smaller component; and
- **preserves the original scale**, since $\mathrm{TU} = \mathrm{OP} = x$ implies $\mathrm{TO} = x$.

Thus, a high TO requires both strong task utility and strong privacy protection.

Comparison with alternative trade-off measures. Table 4 compares several common alternatives. An arithmetic mean permits a high value on one objective to compensate too strongly for a low value on the other. A product penalizes low values, but also systematically contracts the score even when utility and protection are well balanced; for example, $\mathrm{TU} = \mathrm{OP} = 0.8$ gives a product of 0.64, whereas the harmonic mean remains 0.8. The geometric mean alleviates this scale contraction

Algorithm 3: Complementary Expertise Integration (CEI)

Environment : Embedding function $f(\cdot)$.
Input : Protected query $\tilde{q}$; target memories $\mathcal{R}_t(q)$; peer memories $\mathcal{R}_p(\tilde{q})$.
Output : Scored peer evidence $\{(m_p^j, s_p)\}_{j=1}^K$.

// --- Query-Relative Memory Geometry ---
Normalize the protected query:

$$\hat{\mathbf{q}} \leftarrow \frac{f(\tilde{q})}{\|f(\tilde{q})\|}$$

foreach $m \in \mathcal{R}_t(q) \cup \mathcal{R}_p(\tilde{q})$ **do**
  Normalize memory embedding:

$$\hat{\mathbf{m}} \leftarrow \frac{f(m)}{\|f(m)\|}$$

  Compute angular radius:

$$\theta(m) \leftarrow \arccos\left(\hat{\mathbf{q}}^\top \hat{\mathbf{m}}\right)$$

  Compute tangent direction:

$$\mathbf{u}(m) \leftarrow \frac{\hat{\mathbf{m}} - (\hat{\mathbf{q}}^\top \hat{\mathbf{m}})\hat{\mathbf{q}}}{\|\hat{\mathbf{m}} - (\hat{\mathbf{q}}^\top \hat{\mathbf{m}})\hat{\mathbf{q}}\|}$$

// --- Pairwise Complementarity ---
foreach $m_p^j \in \mathcal{R}_p(\tilde{q})$ **do**
  foreach $m_t^i \in \mathcal{R}_t(q)$ **do**
    Compute directional novelty:

$$n_{ji} \leftarrow \frac{\arccos\left(\mathbf{u}(m_p^j)^\top \mathbf{u}(m_t^i)\right)}{\pi}$$

    Compute radial gain:

$$g_{ji} \leftarrow \max\left(0,\ \frac{\theta(m_p^j) - \theta(m_t^i)}{\theta(m_p^j)}\right)$$

Convert $\{n_{ji}\}$ and $\{g_{ji}\}$ to their mid-rank percentiles $\{\bar{n}_{ji}\}$ and $\{\bar{g}_{ji}\}$
foreach *peer–target pair* (j, i) **do**
  Compute joint complementarity:

$$s_{ji} \leftarrow 1 - (1 - \bar{n}_{ji})(1 - \bar{g}_{ji})$$

// --- Personalized Peer Evidence ---
foreach $m_p^j \in \mathcal{R}_p(\tilde{q})$ **do**
  Aggregate its pairwise complementarity:

$$s_p \leftarrow \min_i s_{ji}$$

Rank peer memories according to s_p
return $\{(m_p^j, s_p)\}_{j=1}^K$

but is less sensitive than the harmonic mean to imbalance. Using the minimum of the two objectives imposes an even stricter bottleneck criterion, but completely ignores improvements in the larger component. The harmonic mean therefore provides a balanced compromise between sensitivity to the weaker objective and preservation of the original metric scale.

D Details of Comparison Methods

Counterpart Organization. As discussed in Sec. 2.3, late-stage collaboration involves two communication boundaries: task information sent to peers and reasoning returned to the target. As illustrated in Fig. 1, full-task sharing may expose sensitive information, while fragmentation can remove useful context. On the return side, final answers provide limited collaborative information,

Table 4: Comparison of alternative privacy–utility aggregation functions. U and P denote Task Utility (TU) and Overall Protection (OP), respectively.

Measure	Definition	Scale preserving	Imbalance penalty	Main limitation
Arithmetic mean	$\frac{U+P}{2}$	✓	Weak	Allows strong compensation
Product	UP	×	Strong	Contracts the metric scale
Geometric mean	$\sqrt{UP}$	✓	Moderate	Less sensitive to imbalance
Minimum	$\min(U, P)$	✓	Very strong	Ignores the larger component
Harmonic mean (ours)	$\frac{2UP}{U+P}$	✓	Strong	Aggregate rather than instance-level

whereas explicit evidence may expose private knowledge without fully preserving latent expertise. Accordingly, we organize the counterparts in Table 1 into three groups:

- **Privacy Protecting (without peer):** query-side privacy protection without collaboration.
- **Collaboration (with peer):** peer collaboration without query protection.
- **Privacy Protecting + Collaboration:** joint query protection and peer collaboration.

This organization allows us to separately examine information disclosure and expertise utilization before evaluating their combined effect.

Single Agent [ICLR'23]. We use a non-collaborative single-agent baseline, following the standard LLM-agent paradigm (Yao et al., 2023). The target agent independently reasons over the original query and its locally retrieved memories without peer communication.

MNI-Gate [GAIE'26]. MNI-Gate is the information-filtering component of MIN-Trust (Chen et al., 2026), which operationalizes the Minimum Necessary Information principle for inter-agent communication. It classifies outgoing information into task-essential, contextual, and private content, retaining necessary information while transforming or suppressing less necessary sensitive content before transmission. We adapt MNI-Gate as a query-side privacy counterpart by applying its information-minimization mechanism to the input query without enabling peer collaboration. It represents a gating-based disclosure strategy in which privacy decisions are determined by the necessity of information for the task, in contrast to our agent-specific expertise-loss criterion.

PromptObfus [AAAI'26]. PromptObfus (Li et al., 2026) is a prompt-desensitization method based on anti-adversarial learning. It first identifies privacy-sensitive tokens and replaces them with `[MASK]`, uses a desensitization model to generate candidate substitutions, and selects replacements according to feedback from a surrogate model so that sensitive information is obscured while the original task prediction is minimally affected. We apply PromptObfus to the input query as another query-side privacy counterpart, without peer communication. Compared with MNI-Gate, which filters information according to task necessity, PromptObfus represents token-level privacy obfuscation designed primarily to preserve prediction consistency.

Reasoning Share [ACL'24]. Reasoning Share is a controlled instantiation of the reasoning-level output-sharing paradigm, following the use of expert analyses and reasoning exchange in prior multi-agent systems such as MedAgents (Tang et al., 2024). Each peer receives the unprotected query, retrieves its top-K private memories according to semantic relevance, and generates a reasoning summary for the target agent. The target agent then integrates these peer reasoning signals with its own private evidence for final prediction. Unlike final-answer sharing, this counterpart communicates richer intermediate reasoning; however, it neither protects the query nor explicitly distinguishes complementary peer knowledge from information already available to the target agent.

PrivCollab-Agent [arXiv'26]. PrivCollab-Agent (Jiang et al., 2026) protects inter-agent communication through claim-level information sharing. Rather than transmitting raw private context, each agent communicates structured claims consisting of an abstract conclusion, a confidence estimate, and a minimal evidence tag, which are further controlled by a privacy filter. In our setting, PrivCollab-Agent serves as an output-side privacy counterpart: the original query is shared with peers, while the information returned from peers is restricted through its claim-level communication mechanism. This allows us to evaluate whether directly constraining peer outputs is sufficient for privacy-aware collaboration.

Table 5: Comparison of disclosure criteria. All criteria operate on the same query-local expertise losses and differ only in how the disclosure boundary is determined.

Criterion	Boundary	Preserve Rule	Calibration
Zero Threshold	$\tau = 0$	$x_l > \tau$	Global
Mixture Mean	$\tau = \pi_{\text{low}}\mu_{\text{low}} + \pi_{\text{high}}\mu_{\text{high}}$	$x_l > \tau$	Query-local statistic
Median	$\tau = \text{median}(\{x_l\}_{l=1}^{L})$	$x_l > \tau$	Query-local statistic
GMM Posterior	$\pi_{\text{low}}\mathcal{N}_{\text{low}}(\tau) = \pi_{\text{high}}\mathcal{N}_{\text{high}}(\tau)$	$P(\text{high} \mid x_l) > P(\text{low} \mid x_l)$	Query-local posterior

MNI-Gate + Reasoning Share. This counterpart combines MNI-Gate (Chen et al., 2026) for query-side information minimization with Reasoning Share for peer-to-target reasoning exchange. It represents a straightforward composition of privacy filtering and reasoning-level collaboration, allowing us to examine whether independently combining an existing query-disclosure mechanism with standard reasoning sharing is sufficient to obtain a favorable privacy–utility trade-off.

PromptObfus + Reasoning Share. We similarly combine PromptObfus (Li et al., 2026) with Reasoning Share. Sensitive query information is first desensitized through token replacement before the processed query is sent to peer agents, which subsequently return reasoning summaries to the target agent. This counterpart tests whether prediction-preserving prompt obfuscation can be directly combined with late-stage reasoning exchange.

RAD + PrivCollab-Agent. This counterpart combines our RAD module with the output-side privacy mechanism of PrivCollab-Agent (Jiang et al., 2026). RAD controls which sensitive query information is disclosed to peers, while PrivCollab-Agent restricts the information returned from peers through claim-level communication. It therefore provides a privacy-oriented combination at both communication boundaries, without our complementary-expertise identification mechanism.

E Analysis of Query-wise GMM Calibration

We analyze how different disclosure criteria determine the preserve-or-abstract boundary from query-local expertise losses. RAD uses a query-wise Gaussian mixture model (GMM) to infer this boundary from the loss distribution. We compare alternative criteria and establish the Bayes optimality, self-calibration, and consistency of the GMM posterior rule.

E.1 Compared Disclosure Criteria

Figure 5 compares four disclosure criteria that share the same expertise-loss measure $\Delta E_t(z)$ but differ in how the preserve-or-abstract boundary is determined. Let $x_l = \Delta E_t(z_l)$ for the sensitive spans $\{z_l\}_{l=1}^{L}$. Table 5 summarizes these criteria.

E.2 Posterior Decision Rule

For a given query, let

$$x = \Delta E_t(z)$$

denote the expertise loss associated with a sensitive span z. We model the query-local expertise-loss distribution using a two-component Gaussian mixture:

$$p(x) = \pi_{\text{low}}\mathcal{N}\left(x; \mu_{\text{low}}, \sigma^2_{\text{low}}\right) + \pi_{\text{high}}\mathcal{N}\left(x; \mu_{\text{high}}, \sigma^2_{\text{high}}\right), \tag{23}$$

where

$$\mu_{\text{low}} < \mu_{\text{high}}, \qquad \pi_{\text{low}} + \pi_{\text{high}} = 1.$$

The two components respectively represent relatively low and high expertise-loss spans within the current query. The posterior assignment is

$$\hat{c}(x) = \arg\max_{c \in \{\text{low},\text{high}\}} \pi_c \mathcal{N}\left(x; \mu_c, \sigma_c^2\right). \tag{24}$$

Accordingly, spans assigned to the high-loss component are preserved, whereas spans assigned to the low-loss component are abstracted.

A decision boundary τ is therefore a point at which the two posterior scores are equal:

$$\pi_{\text{low}}\mathcal{N}\left(\tau;\mu_{\text{low}},\sigma^2_{\text{low}}\right)=\pi_{\text{high}}\mathcal{N}\left(\tau;\mu_{\text{high}},\sigma^2_{\text{high}}\right). \tag{25}$$

At this point, the two components are equally plausible under the fitted query-specific mixture.

E.3 Bayes Optimality

Proposition 1. For a fixed fitted mixture in eq. (23), the posterior rule in eq. (24) minimizes the conditional classification error between the low- and high-loss components.

Proof. For an observed expertise loss x, assigning it to the low-loss component incurs conditional error

$$P(c=\text{high}\mid x),$$

whereas assigning it to the high-loss component incurs

$$P(c=\text{low}\mid x).$$

Therefore, the minimum conditional error is achieved by selecting the component with the larger posterior probability:

$$\hat{c}(x)=\arg\max_c P(c\mid x).$$

By Bayes' rule,

$$P(c\mid x)\propto\pi_c\mathcal{N}(x;\mu_c,\sigma_c^2),$$

which is exactly the rule in eq. (24). Thus, the GMM assignment is Bayes optimal under the fitted query-local mixture. □

This result shows that the GMM does not introduce an arbitrary disclosure threshold. Instead, its boundary is the Bayes decision boundary induced by the estimated expertise-loss distribution of the current query.

E.4 Existence of an Interior Calibration Boundary

We next show when the GMM induces a unique separating boundary between the two component means.

Define the log posterior odds

$$G(x)=\log\frac{\pi_{\text{high}}\mathcal{N}\left(x;\mu_{\text{high}},\sigma^2_{\text{high}}\right)}{\pi_{\text{low}}\mathcal{N}\left(x;\mu_{\text{low}},\sigma^2_{\text{low}}\right)}. \tag{26}$$

The decision boundary satisfies $G(\tau)=0$. Let

$$\Delta\mu=\mu_{\text{high}}-\mu_{\text{low}}>0$$

and

$$A=\log\frac{\pi_{\text{high}}\sigma_{\text{low}}}{\pi_{\text{low}}\sigma_{\text{high}}}.$$

Evaluating eq. (26) at the two component means gives

$$G(\mu_{\text{low}})=A-\frac{\Delta\mu^2}{2\sigma^2_{\text{high}}}, \tag{27}$$

and

$$G(\mu_{\text{high}})=A+\frac{\Delta\mu^2}{2\sigma^2_{\text{low}}}. \tag{28}$$

Proposition 2. If

$$-\frac{\Delta\mu^2}{2\sigma^2_{\text{low}}}<\log\frac{\pi_{\text{high}}\sigma_{\text{low}}}{\pi_{\text{low}}\sigma_{\text{high}}}<\frac{\Delta\mu^2}{2\sigma^2_{\text{high}}}, \tag{29}$$

then there exists a unique decision boundary

$$\tau\in(\mu_{\text{low}},\mu_{\text{high}}).$$

Proof. Under eq. (29), eqs. (27) and (28) imply

$$G(\mu_{\text{low}}) < 0, \qquad G(\mu_{\text{high}}) > 0.$$

Since $G(x)$ is continuous, the intermediate value theorem guarantees at least one root

$$\tau \in (\mu_{\text{low}}, \mu_{\text{high}}).$$

It remains to show uniqueness within this interval. Differentiating eq. (26) gives

$$G'(x) = \frac{x - \mu_{\text{low}}}{\sigma_{\text{low}}^2} - \frac{x - \mu_{\text{high}}}{\sigma_{\text{high}}^2}. \tag{30}$$

At the two endpoints,

$$G'(\mu_{\text{low}}) = \frac{\Delta\mu}{\sigma_{\text{high}}^2} > 0,$$

and

$$G'(\mu_{\text{high}}) = \frac{\Delta\mu}{\sigma_{\text{low}}^2} > 0.$$

Because $G'(x)$ is linear in x, positivity at both endpoints implies

$$G'(x) > 0, \qquad x \in [\mu_{\text{low}}, \mu_{\text{high}}].$$

Hence, $G(x)$ is strictly increasing between the two component means and can cross zero only once. Therefore, the boundary in $(\mu_{\text{low}}, \mu_{\text{high}})$ is unique. □

The condition in eq. (29) has a natural interpretation: the separation between the two component means must be sufficiently strong relative to the imbalance in their mixture weights and variances. When this holds, the query-wise GMM yields a well-defined boundary that separates relatively low- and high-expertise-loss spans.

E.5 Closed-form Boundary under Equal Variance

The self-calibrating behavior becomes particularly transparent when the two components have the same variance,

$$\sigma_{\text{low}} = \sigma_{\text{high}} = \sigma.$$

Corollary 1. Under equal component variance, the decision boundary has the closed form

$$\tau = \frac{\mu_{\text{low}} + \mu_{\text{high}}}{2} + \frac{\sigma^2}{\mu_{\text{high}} - \mu_{\text{low}}} \log \frac{\pi_{\text{low}}}{\pi_{\text{high}}}. \tag{31}$$

Proof. Substituting the equal-variance assumption into eq. (25) and taking logarithms gives

$$\log \pi_{\text{low}} - \frac{(\tau - \mu_{\text{low}})^2}{2\sigma^2} = \log \pi_{\text{high}} - \frac{(\tau - \mu_{\text{high}})^2}{2\sigma^2}.$$

Rearranging yields eq. (31). □

When the two components have equal prior weights,

$$\pi_{\text{low}} = \pi_{\text{high}},$$

the correction term vanishes, and

$$\tau = \frac{\mu_{\text{low}} + \mu_{\text{high}}}{2}. \tag{32}$$

Thus, in the balanced case, the GMM boundary reduces exactly to the midpoint between the two query-local clusters. For imbalanced mixtures, the boundary is automatically shifted according to their estimated relative prevalence rather than remaining fixed at the midpoint.

E.6 Query-wise Self-Calibration

A central motivation for using a query-wise mixture model is that expertise losses can exhibit different locations and scales across queries. A globally fixed threshold does not adapt to such variation. The GMM posterior rule, in contrast, is invariant to positive affine transformations.

Proposition 3. Consider a transformed expertise-loss variable

$$x' = ax + b, \qquad a > 0.$$

Let the corresponding mixture parameters be

$$\mu'_c = a\mu_c + b, \qquad \sigma'_c = a\sigma_c, \qquad \pi'_c = \pi_c.$$

Then the posterior assignment is unchanged:

$$\hat{c}(x') = \hat{c}(x). \tag{33}$$

Moreover, if τ is a decision boundary before transformation, the corresponding transformed boundary is

$$\tau' = a\tau + b. \tag{34}$$

Proof. For each component,

$$\mathcal{N}\left(x'; \mu'_c, {\sigma'_c}^2\right) = \frac{1}{a}\mathcal{N}\left(x; \mu_c, \sigma_c^2\right). \tag{35}$$

Therefore,

$$\frac{\pi_{\text{high}}\mathcal{N}\left(x'; \mu'_{\text{high}}, {\sigma'_{\text{high}}}^2\right)}{\pi_{\text{low}}\mathcal{N}\left(x'; \mu'_{\text{low}}, {\sigma'_{\text{low}}}^2\right)} = \frac{\pi_{\text{high}}\mathcal{N}\left(x; \mu_{\text{high}}, \sigma^2_{\text{high}}\right)}{\pi_{\text{low}}\mathcal{N}\left(x; \mu_{\text{low}}, \sigma^2_{\text{low}}\right)},$$

since the common factor $1/a$ cancels. Hence, the posterior ordering between the two components is preserved, which proves eq. (33). Applying the same transformation to any point satisfying eq. (25) immediately yields eq. (34). □

This invariance provides a precise interpretation of *query-wise self-calibration*. The disclosure decision depends on the relative structure of expertise losses within the current query rather than on their absolute numerical scale. Consequently, a query whose expertise-loss distribution is globally shifted or rescaled induces the same relative preserve-or-abstract decisions, whereas a globally fixed threshold generally does not preserve this property.

E.7 Consistency of the Estimated Boundary

Finally, we characterize the asymptotic behavior of the estimated GMM boundary. Let the population expertise-loss distribution be generated by a two-component Gaussian mixture with true parameters

$$\theta^\star = \{\pi_c^\star, \mu_c^\star, \sigma_c^\star\}_{c\in\{\text{low,high}\}},$$

where

$$\mu^\star_{\text{low}} < \mu^\star_{\text{high}}.$$

Let $\hat{\theta}_L$ denote the maximum-likelihood estimate obtained from L query-local expertise-loss observations, and let $\hat{\tau}_L$ denote the corresponding interior decision boundary.

Proposition 4. Assume that the two-component Gaussian mixture is identifiable, its parameters lie in a regular non-degenerate parameter region, and the population decision boundary $\tau^\star$ is an isolated root satisfying

$$G_{\theta^\star}(\tau^\star) = 0, \qquad G'_{\theta^\star}(\tau^\star) \neq 0.$$

If

$$\hat{\theta}_L \xrightarrow{p} \theta^\star,$$

then

$$\hat{\tau}_L \xrightarrow{p} \tau^\star. \tag{36}$$

Table 6: Effect of the number of agents on MedQA with Qwen3-30B.

# Agents	Method	TU(↑)	QP(↑)	MP(↑)	OP(↑)	TO(↑)
3	MNI-Gate + Reasoning Share	81.23	62.11	97.71	94.91	87.54
	PromptObfus + Reasoning Share	81.23	45.87	97.35	93.57	86.96
	Ours	**84.79**	50.21	96.37	93.03	**88.72**
4	MNI-Gate + Reasoning Share	77.91	65.22	97.02	94.42	85.38
	PromptObfus + Reasoning Share	80.58	45.87	96.61	92.85	86.28
	Ours	**83.25**	51.34	95.71	92.41	**87.59**
5	MNI-Gate + Reasoning Share	76.89	64.97	97.00	95.04	85.01
	PromptObfus + Reasoning Share	80.97	45.87	96.92	94.03	87.01
	Ours	**83.50**	50.80	96.00	93.44	**88.19**

Proof. The log posterior odds

$$G_\theta(x)$$

is continuous in both x and the mixture parameters whenever the component variances are positive. Therefore,

$$\hat{\theta}_L \xrightarrow{p} \theta^\star$$

implies that

$$G_{\hat{\theta}_L}(x) \xrightarrow{p} G_{\theta^\star}(x)$$

locally around $\tau^\star$. Because $\tau^\star$ is an isolated simple root with

$$G'_{\theta^\star}(\tau^\star) \neq 0,$$

standard root-continuity arguments imply that the corresponding estimated root converges to the population root:

$$\hat{\tau}_L \xrightarrow{p} \tau^\star.$$

Therefore, when the query-local expertise-loss distribution exhibits a well-defined two-component structure, the GMM boundary approaches the population Bayes boundary as more observations become available. Conversely, when the observations are insufficient or the two components cannot be reliably distinguished, such a calibrated separation is not supported by the data. This motivates the privacy-first fallback adopted by RAD, which abstracts all detected sensitive spans in degenerate or non-separable cases.

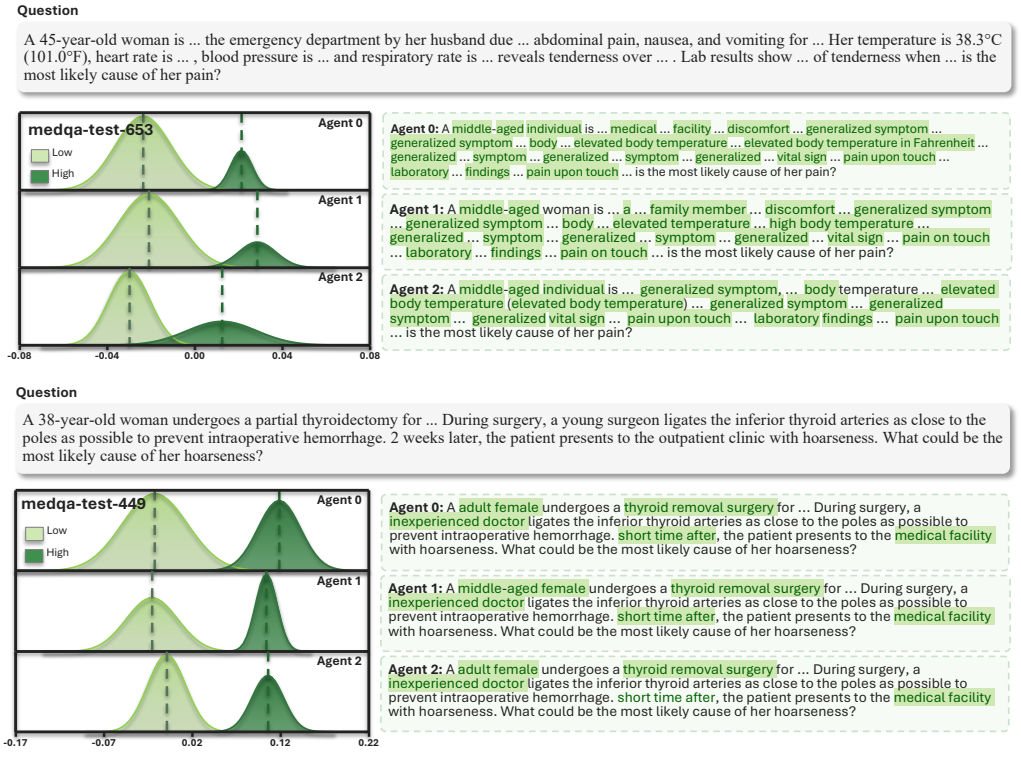
E.8 Qualitative Analysis of Adaptive Disclosure

As shown in Fig. 9, we visualize RAD on MedQA and LegalQA. For each query, the left panels show the query-local GMM over expertise losses, while the right panels show the resulting privacy-controlled queries for different target agents.

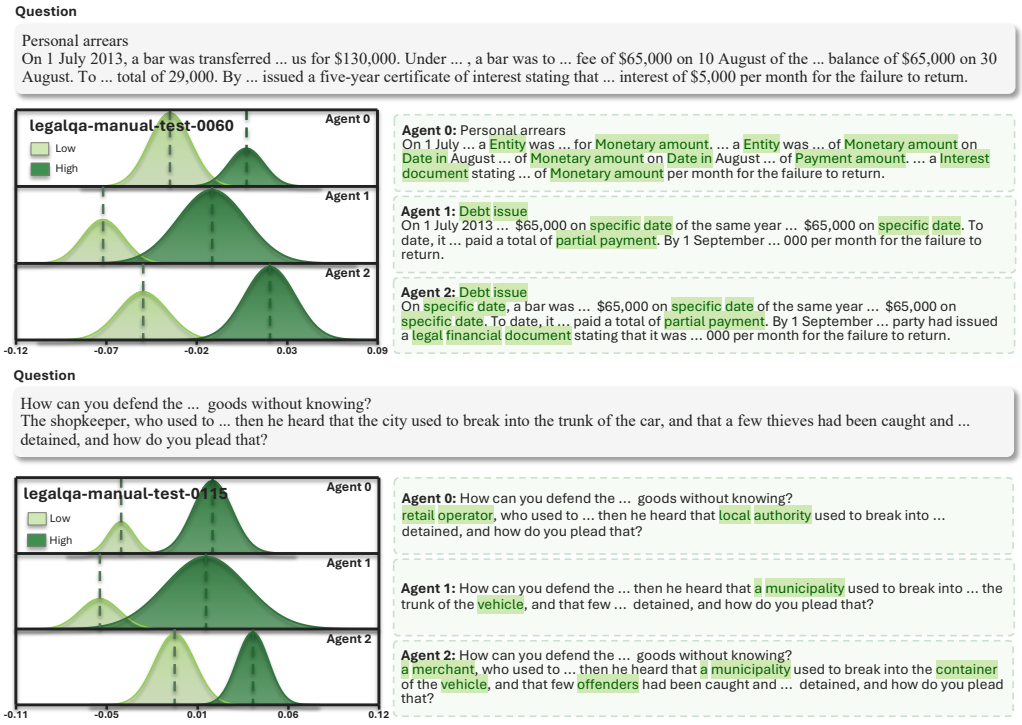
The same query can induce different disclosure decisions because sensitive spans have different expertise relevance under heterogeneous private memories. Accordingly, high-loss spans are preserved, whereas low-loss spans are abstracted. Across both medical and legal examples, RAD selectively retains task-relevant details while generalizing less informative sensitive content, demonstrating agent- and query-specific disclosure rather than uniform masking.

F Effect of the Number of Agents.

As shown in Table 6, Ours consistently achieves the highest TU and TO across different numbers of agents, indicating stable collaborative performance as the system scales from 3 to 5 agents. In particular, the 3-agent setting attains the best TU (84.79) and TO (88.72) for Ours. Increasing the number of agents does not yield consistent gains: the performance of MNI-Gate + Reasoning Share decreases noticeably, while PromptObfus + Reasoning Share shows only minor fluctuations. These results suggest that additional agents do not necessarily provide more useful complementary expertise. We therefore adopt 3 agents as the default setting in our experiments.



(a) Qualitative examples on MedQA.



(b) Qualitative examples on LegalQA.

Figure 9: Qualitative visualization of query-local adaptive disclosure on MedQA and LegalQA. For each query, the left panels show the fitted low- and high-expertise-loss components for different target agents, while the right panels show the resulting privacy-controlled queries.