PFEDMA: WHO SHOULD JOIN NEXT AS FEDERATED MULTI-AGENT ORCHESTRATION EVOLVES?

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated Multi-Agent Systems (FedMAS) offer a promising paradigm for collaborative reasoning among LLM-based agents while keeping raw client data decentralized. However, reliable multi-round collaboration remains challenging. Agent selection must take into consideration evolving task requirements and the agent's historical reliability, while privacy rules must adapt to changing collaboration contexts and meet the client's personalized needs. To address these challenges, we propose **PFedMA**, a federated multi-agent framework for feedback driven multi-round collaboration under heterogeneous privacy requirements. Across rounds, PFedMA maintains historical capability and reliability states to guide agent selection, evaluates task responses before aggregation, and refines subsequent queries according to accumulated collaboration feedback. PFedMA enables bidirectional local to global and global to local private rule evolution across heterogeneous clients. Clients generalize locally updated privacy rules and selectively incorporate globally maintained rules, thereby facilitating cross-client privacy rule evolution while preserving personalized constraints. Extensive experiments across reasoning, privacy sensitive, and adversarial settings show that PFedMA improves task performance, reduces measured privacy leakage, and improves system robustness. *The implementation code for PFedMA is available at* `https://anonymous.4open.science/r/paper_PFedMA-2F44/`.

1 INTRODUCTION

LLM based multi-agent systems (MAS) coordinate agents with complementary capabilities to address complex tasks (Guo et al., 2024; Wang et al., 2024). In privacy sensitive settings, however, the knowledge required for collaboration is distributed across multiple data holders with distinct privacy requirements, data ownership constraints, and access control policies. To complete the collaborative task, agents must integrate distributed knowledge and intermediate reasoning information, while raw data cannot be centralized. This challenge motivates federated multi-agent systems (FedMAS), a distributed collaboration paradigm in which agents reason over locally held information and exchange task relevant information without directly centralizing raw client data (Shi et al., 2025). However, existing FedMAS does not fully address how to select reliable agents or adapt information sharing rules as reasoning unfolds across rounds.

As collaboration progresses, both *who should participate* and *what information should be shared* may change with the evolving reasoning state. Early rounds may require broad domain expertise, whereas later rounds may call for specialized agents to resolve uncertainties exposed by intermediate responses. Meanwhile, as reasoning progresses and collaboration demands evolve, newly required information may introduce sharing scenarios that are not covered by existing privacy rules. Consequently, agent orchestration and privacy control become closely interdependent across reasoning rounds. *Agent orchestration determines whose knowledge enters the collaborative reasoning process*, while *privacy rules determine what information can be shared*. The resulting responses and reasoning state, in turn, affect both subsequent agent selection and the privacy constraints relevant to later rounds. FedMAS therefore needs to use feedback from intermediate responses to update both agent selection and privacy rules across rounds.

These two interdependent decisions give rise to two challenges. First, task response reliability must inform agent selection across rounds. Previous studies have explored iterative interactions between

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.

clients in federated collaboration (Wang et al., 2025), as well as multi-round routing based on the current interaction context (Zhang et al., 2025a). However, selecting agents based only on the current query context may overlook agents' historical reliability in previous rounds. Because agents may produce incomplete, erroneous, or adversarial responses (Deng et al., 2025; Zhang & Yang, 2026), selecting agents without accounting for their past reliability can propagate errors across rounds. Second, privacy rules must adapt to evolving collaboration while preserving heterogeneous client requirements. Existing agent-level approaches protect sensitive information through context filtering and rule-based protection (Shi et al., 2025; Xiang et al., 2025). Separately, applying DP to agent responses (Panda et al., 2023) may introduce perturbations that propagate across rounds, increasing the risk of semantic drift and inconsistent reasoning. These approaches typically rely on predefined privacy constraints or fixed protection strategies, which may fail to accommodate changing information needs across reasoning rounds, either suppressing information necessary for subsequent reasoning or failing to capture newly emerging privacy risks (Truong et al., 2021; Shi et al., 2025). Moreover updating rules only locally limits cross client reuse, while applying shared rules uniformly can overlook differences in client requirements, data characteristics, and task contexts (Kossek & Stefanovic, 2024). The challenge is to reuse newly identified privacy rules across clients without overriding their distinct local requirements.

To address these challenges, we propose **PFedMA**, a federated multi-agent framework for feedback driven multi-round collaboration under heterogeneous privacy requirements. At each round, PFedMA evaluates task responses and updates historical capability and reliability estimates to guide agent selection and query refinement in the next round. Response reliability also determines which clients contribute new privacy rule candidates. These candidates are generalized for cross-client reuse, while each client selectively adopts rules according to its local requirements; the adopted rules then govern subsequent queries and responses.

Our main contributions are summarized as follows:

- We propose PFedMA, a federated multi-agent system with feedback driven orchestration that leverages cross round capability and reliability feedback to adapt agent selection, thereby mitigating repeated selection of unreliable agents and error accumulation across rounds.
- We introduce a federated privacy rule co-evolution mechanism in PFedMA that enables privacy rules to adapt to evolving collaboration while preserving heterogeneous client requirements, allowing local privacy rules to benefit other clients without overriding personalized privacy rules.
- Experiments across diverse benchmarks show that PFedMA improves average task performance and reduces measured privacy leakage compared with Fed-ICL, while achieving a lower attack success rate than EPEAgents under the evaluated prompt injection attacks.

2 Related Work

Multi-Agent System. Multi-agent systems (MAS) predominantly rely on centralized architectures, structuring collaboration via domain personas (Tang et al., 2024), external retrieval (Xiao et al., 2025), or predefined execution pipelines (Chen et al., 2023; Wang et al., 2024). To resolve conflicts and reach consensus, systems employ diverse alignment strategies, including textual debates and voting (Choi et al., 2025; Sun et al., 2025), adaptive termination (Hu et al., 2025), and belief state broadcasting (Deng et al., 2025; Keshmirian et al., 2025). Recent advances in task routing further optimize this orchestration using dynamic paths (Yue et al., 2025) and reinforcement learning (Zhang et al., 2025a). However, these centralized paradigms mandate transparent raw context sharing, inherently violating data privacy. In contrast, our framework achieves dynamic multi-round dispatch while operating under explicit local data boundaries.

Federated Multi-Agent System. To reduce communication overhead, federated paradigms increasingly leverage text driven prompting (Qiu et al., 2023), textual gradients (Chen et al., 2025), and in context learning (Wu et al., 2024). Similarly, cooperative frameworks iteratively aggregate plain text responses from local exemplars (Wang et al., 2025). However, sharing raw text introduces severe prompt extraction risks (Duan et al., 2024), making secure communication a critical challenge (Kossek & Stefanovic, 2024). Existing privacy enhancing approaches, including trusted environments (Nagar et al., 2021), node decomposition (Wang et al., 2021), and differential privacy (Panda et al., 2023), reduce information exposure through different mechanisms. However,

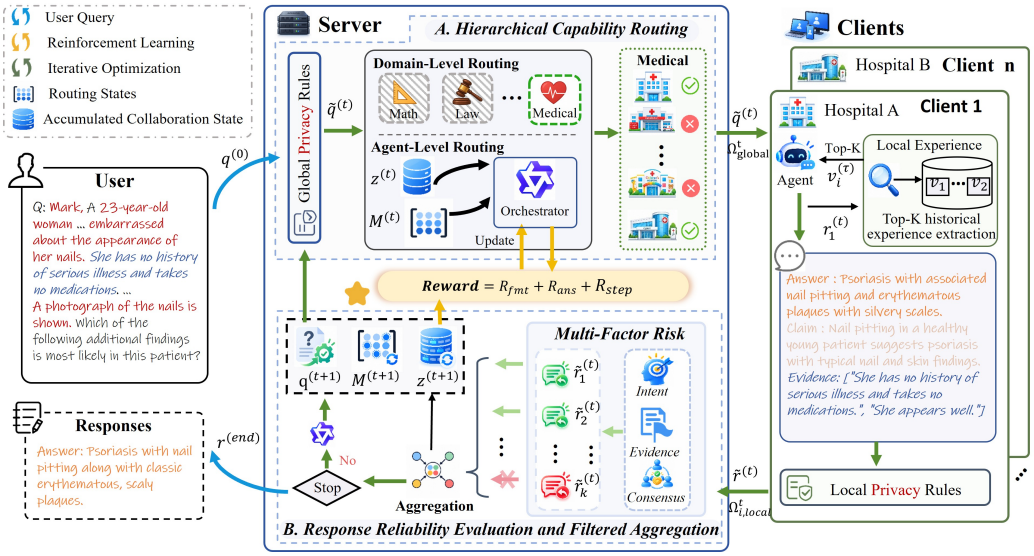


Figure 1: Overview of PFedMA. Task response feedback guides query refinement and client routing, while privacy rules evolve through cross client sharing and personalized adaptation. Together, they jointly determine the participants and the content to be shared, thereby laying the foundation for continued collaboration. The algorithm is provided in Appendix A.

depending on the mechanism, they may require additional trusted infrastructure or restrict the information available during collaboration, making it difficult to preserve semantically useful context across multiple reasoning rounds (Shi et al., 2025). To mitigate this risk, we restrict server bound communication to privacy processed responses, generalized privacy rule candidates and compact rule level feedback, thereby preserving the stated client local data boundary.

3 METHOD

3.1 PRELIMINARIES

We consider a federated multi-agent system with an honest but curious server and N heterogeneous clients $\mathcal{A} = \{A_1, \ldots, A_N\}$. Each client A_i maintains a private dataset $\mathcal{D}_i$ and a local experience buffer $\mathcal{X}_i$, both of which remain on device. Although the server cannot directly access clients' raw local data or unprocessed local responses, it may attempt to infer private information from uploaded privacy processed responses $\tilde{r}_i^t$. Meanwhile, some clients may provide unreliable or malicious task responses.

To address these concerns, we introduce historical capability and reliability states, along with global and personalized privacy rule repositories for query and response privacy processing, respectively. At round t, the server maintains the accumulated collaboration state z^{t-1} and processes the query q^t using its global privacy rule repository Ω^t_{global} to obtain $\tilde{q}^t$. The orchestrator π_θ selects a subset of participating clients based on their routing state M_i^t and sends $\tilde{q}^t$ to the selected clients. Each selected client generates a local response $r_i^t = A_i(\tilde{q}^t, \mathcal{D}_i, \mathcal{X}_i)$ and processes it using its personalized privacy rule repository $\Omega^t_{i,\text{local}}$ before uploading $\tilde{r}_i^t$. Both r_i^t and $\Omega^t_{i,\text{local}}$ remain on device. The server evaluates the uploaded responses, updates the routing states to M_i^{t+1}, and aggregates reliable responses to obtain z^t. Meanwhile, global and personalized privacy rules co-evolve through generalized local rule sharing and selective global to local adaptation. If the task remains unresolved, the server refines the query using the updated routing states and applies the updated global privacy rules before dispatching it to selected clients in the next round. The process ends when no further clients are needed or the round budget is exhausted.

3.2 Iterative Orchestration for Federated MAS

Building on the overview in Section 3.1, this section details client selection, response evaluation and aggregation, and the reward used to train the orchestrator.

3.2.1 Agent Selection and Routing State Updates

PFedMA first identifies the expertise domain ρ_d most relevant to $\tilde{q}^t$ as $d_\star^t = \arg\max_{d \in \mathcal{D}omain} \mathrm{Cos}(\tilde{q}^t, \rho_d)$, and then retrieves the corresponding candidate set $\mathcal{A}_{d_\star^t}$. The orchestrator then selects participating agents as

$$\mathcal{E}^t = \pi_\theta\Big(\mathcal{A}_{d_\star^t} \mid \tilde{q}^t, z^{t-1}, \{M_i^t\}_{A_i \in \mathcal{A}_{d_\star^t}}\Big). \tag{1}$$

Where $M_i^t = [\mathbf{C}_i^t \| S_i^t]$ maintains the historical capability and reliability states of client A_i. After the returned response of A_i is evaluated, its routing state is updated using the newly observed historical capability and reliability feedback

$$\mathbf{C}_i^{t+1} = \mathbf{C}_i^t + f_i^t \bar{\mathbf{v}}_i^t, \qquad S_i^{t+1} = \frac{n_i^t S_i^t + [\mathrm{Int}_i^t, \mathrm{Evd}_i^t, \mathrm{Con}_i^t]^\top}{n_i^t + 1}, \qquad A_i \in \widehat{\mathcal{E}}^t \tag{2}$$

where $\widehat{\mathcal{E}}^t$ denotes the set of responses retained by the reliability filter, n_i^t is the number of previously evaluated responses from A_i, and Int_i^t, Evd_i^t, and Con_i^t denote its anomaly scores. We define $\mathbf{C}_i^t$ as accumulated historical capability state of client A_i, where f_i^t indicates whether the current retained response improves upon the most recent retained response of the same client, and $\bar{\mathbf{v}}_i^t$ identifies the capability dimensions in which the current response improves upon that reference. In contrast, S_i^t summarizes the agent's historical risk as the mean of its observed anomaly scores.

The updated routing state is $M_i^{t+1} = \left[\mathbf{C}_i^{t+1} \| S_i^{t+1}\right]$. Thus, response evaluations from the current round are retained in M_i^{t+1} and directly influence agent selection in subsequent rounds. Detailed state update definitions are provided in Appendix D.

3.2.2 Response Reliability Evaluation and Filtered Aggregation

Similar to the recently proposed decision centered model Jev (Almeida, 2026), PFedMA explicitly presents task level results rather than burying them in free form responses. While Jev focuses on decision making, PFedMA focuses on conclusions and records the claims and evidence corresponding to each conclusion. At round t, the privacy processed response of client A_i is represented as $\tilde{r}_i^t = \langle a_i^t, \Gamma_i^t, \mathcal{L}_i^t \rangle$, where a_i^t denotes the final conclusion, Γ_i^t denotes the collection of claims supporting the conclusion, and $\mathcal{L}_i^t$ denotes the corresponding collection of evidence associated with these claims. The details are provided in Appendix C.1.

PFedMA evaluates each response from the complementary perspectives of intent, evidence, and cross agent consensus. To aggregate fine grained anomaly signals, we use

$$\mathrm{Pool}_{\lambda_{pool}}(\{x_j\}_{j=1}^J) = \lambda_{pool} \max_j x_j + (1 - \lambda_{pool}) \frac{1}{J} \sum_{j=1}^J x_j, \tag{3}$$

where J denotes the number of fine grained anomaly scores being aggregated and x_j is the j-th score. The operator balances the most severe anomaly with the overall anomaly level. This score captures abnormal intents such as instruction overriding, false authority claims, aggregator manipulation, and task goal substitution. The evidence anomaly Evd_i^t assesses claim level evidence quality from support, verifiability, and coherence, while the consensus anomaly Con_i^t measures semantic deviation from peer conclusions. Intent and evidence are evaluated by a frozen judge model, whereas consensus is computed deterministically; details are provided in Appendix C.3 and Appendix C.2.

The three anomaly dimensions are combined into an overall anomaly score, and task responses exceeding the tolerance threshold are excluded:

$$\mathcal{R}isk_i^t = \mathrm{Int}_i^t + \mathrm{Evd}_i^t + \alpha_{\mathrm{con}} \mathrm{Con}_i^t, \qquad \widehat{\mathcal{E}}^t = \left\{ A_i \in \mathcal{E}^t \mid \mathcal{R}isk_i^t \le \tau_{\mathrm{risk}} \right\}. \tag{4}$$

where α_{con} controls the contribution of consensus anomaly, and τ_{risk} denotes the anomaly tolerance threshold. A relatively small α_{con} is used because disagreement alone does not necessarily indicate an unreliable response. The retained set $\widehat{\mathcal{E}}^t$ is then used for aggregation.

The retained privacy processed responses are aggregated to update the accumulated collaboration state:

$$z^t = \begin{cases} \text{Fusion}\left(\{\tilde{r}_i^t\}_{A_i \in \widehat{\mathcal{E}}^t}, z^{t-1}\right), & \widehat{\mathcal{E}}^t \neq \emptyset, \\ z^{t-1}, & \widehat{\mathcal{E}}^t = \emptyset. \end{cases} \tag{5}$$

In our implementation, Fusion is realized by the server side orchestrator π_θ, which synthesizes the retained privacy processed responses and the previous global reasoning state into z^t.

3.2.3 Composite Reward for Multi-Round Orchestration

During RL training, PFedMA uses a composite reward that balances final answer correctness, answer quality improvements across rounds, response reliability, and agent invocation cost. For a collaboration trajectory τ containing T routing rounds, the total reward is

$$R(\tau) = R_{\text{fmt}}(\tau) + R_{\text{ans}}(\tau) + \sum_{t=1}^{T} R_{\text{step}}^t, \tag{6}$$

where R_{fmt} evaluates format compliance and R_{ans} evaluates the final answer. Details for these two formulas are provided in Appendix E.1. Meanwhile, R_{step}^t delivers intermediate feedback for routing, with the step wise reward defined as

$$R_{\text{step}}^t = \Delta Q_t - \overline{\mathcal{R}isk}_t - \frac{|\mathcal{E}^t|}{K_{\max}}, \qquad \overline{\mathcal{R}isk}_t = \frac{1}{|\mathcal{E}^t|} \sum_{A_i \in \mathcal{E}^t} \frac{\mathcal{R}isk_i^t}{2 + \alpha_{\text{con}}}, \qquad |\mathcal{E}^t| > 1. \tag{7}$$

Here, $\Delta Q_t = Q_t - Q_{t-1}$ records the change in answer quality from z^{t-1} to z^t, while $\overline{\mathcal{R}isk}_t$ measures the average normalized response risk of the agents invoked at round t. $K_{\max}$ denotes the maximum number of agents that can be invoked in a single round, and $|\mathcal{E}^t|/K_{\max}$ represents the fraction of the invocation budget used by the current routing decision. Detailed definitions and RL settings are provided in Appendix E.1.

3.3 Federated Privacy Rule Co-Evolution Mechanism

Existing privacy rule based data minimization approaches for Federated MAS mainly regulate information disclosure according to agent roles and access requirements (Shi et al., 2025), but do not address how client specific privacy knowledge can be reused across the federation without centralizing personalized policies. To bridge this gap, PFedMA introduces a *federated privacy rule co-evolution* mechanism, in which shareable privacy knowledge is propagated globally while complete client specific privacy rule repositories remain local.

Global Privacy Rule Repository Ω_{global}^t. The global privacy rule repository is maintained by the server to preserve privacy knowledge that is shareable across clients. At round t, the server receives generalized rule candidates $\bar{\omega}_i^t$ from clients and evaluates their cross client semantic support. When multiple candidates are available, candidates achieving the highest positive support proportion are selected to form the global candidate set:

$$\mathcal{G}^t = \left\{ \bar{\omega}_i^t \in \mathcal{B}^t \;\middle|\; \zeta_i^t > 0,\ \zeta_i^t = \max_{\bar{\omega}_j^t \in \mathcal{B}^t} \zeta_j^t \right\}, \qquad |\mathcal{B}^t| > 1. \tag{8}$$

Here, $\mathcal{B}^t$ is the candidate pool, and $\zeta_i^t = \text{Supp}^t(\bar{\omega}_i^t)/(|\mathcal{B}^t| - 1)$ is the support proportion of candidate $\bar{\omega}_i^t$. The set $\mathcal{G}^t$ retains all candidates with the highest positive support proportion. Details are provided in Appendix F. The server also collects compact rule level feedback $\mathcal{F}^t$ from previous rule executions and uses it together with the promoted candidates to update the global repository:

$$\Omega_{\text{global}}^{t+1} = \text{RuleUpdate}\left(\Omega_{\text{global}}^t, \mathcal{G}^t, \mathcal{F}^t\right). \tag{9}$$

This update consolidates shareable rules and resolves conflicting or ineffective global rules according to accumulated client feedback. The updated repository is then used by the server to process

the next query before dispatch: $\tilde{q}^{t+1} = P_{\text{in}}\left(q^{t+1}, \Omega^{t+1}_{\text{global}}\right)$. The resulting privacy rule processed query $\tilde{q}^{t+1}$ is then dispatched to the selected clients for local inference, thereby transferring the updated global privacy constraints to the next round of collaboration. Detailed definitions and update procedures are provided in Appendix F.

Personalized Privacy Rule Repository $\Omega^t_{i,\text{local}}$. Upon receiving the privacy processed query $\tilde{q}^{t+1}$, client A_i performs local inference using its private data and accumulated experience $r_i^{t+1} = A_i\left(\tilde{q}^{t+1}, \mathcal{D}_i, \mathcal{X}_i\right)$. $\Phi_i^t(\omega)$ denotes the local utility score of privacy rule ω for client A_i at communication round t, and is defined as follows:

$$\Phi_i^t(\omega) = \text{Rel}\left(\omega, \mathcal{X}_i\right) + \frac{n^t_{i,\text{valid}}(\omega)}{\max\left(1, n^t_{i,\text{trigger}}(\omega)\right)} - \frac{n^t_{i,\text{conflict}}(\omega)}{\max\left(1, n^t_{i,\text{trigger}}(\omega)\right)} \tag{10}$$

Here, the term $\text{Rel}(\omega, \mathcal{X}_i)$ measures the semantic relevance of rule ω to the local experience buffer $\mathcal{X}_i$. Furthermore, $n^t_{i,\text{trigger}}(\omega)$, $n^t_{i,\text{valid}}(\omega)$, and $n^t_{i,\text{conflict}}(\omega)$ denote the numbers of times that rule ω is triggered, successfully enforced, and found to conflict with a local privacy requirement, respectively. The details are provided in Appendix F.

Meanwhile, the updated global rules are redistributed to the client for personalized adaptation. Rather than directly replacing its local rules, client A_i compares each redistributed global rule ω_g^{t+1} with its semantically corresponding local rule ω_i^t according to their local utility and selects

$$\omega_i^\star = \begin{cases} \omega_i^t, & \Phi_i^t(\omega_i^t) \geq \Phi_i^t(\omega_g^{t+1}), \\ \omega_g^{t+1}, & \text{otherwise}, \end{cases} \qquad \Omega^{t+1}_{i,\text{local}} = \Omega^t_{i,\text{local}} \cup_{merge} \{\omega_i^\star\}. \tag{11}$$

Here, $\omega_g^{t+1} \in \Omega^{t+1}_{\text{global}}$, with a detailed definition provided in Appendix F. This selective adaptation allows global privacy knowledge to refine the personalized repository without overriding client specific alternatives. The personalized repository is applied to each locally generated response before transmission $\tilde{r}_i^{t+1} = P_{\text{resp},i}\left(r_i^{t+1}, \Omega^{t+1}_{i,\text{local}}\right)$. If a previously uncovered privacy constraint is detected, the corresponding rule $\omega_i^{t+1} \in \Omega^{t+1}_{i,\text{local}}$ remains on device, and only its generalized representation is shared $\bar{\omega}_i^{t+1} = \mathcal{P}_i^{\text{gen}}\left(\omega_i^{t+1}\right)$. Here, $\mathcal{P}_i^{\text{gen}}$ denotes the client side private rules generalization operator, and $\bar{\omega}_i^{t+1}$ is the shareable rule representation. The generalized rule is sent to the server for the subsequent global privacy rule updates; see Appendix F.

4 Experiments

In the experiments, we seek to answer the following research questions:

RQ1: Does cross round response feedback improve multi-round federated reasoning over feedback free orchestration?

RQ2: Can federated privacy rule co-evolution enable cross client privacy rule coordination and personalized adaptation while maintaining rule enforcement effectiveness and task utility?

RQ3: Can multi-factor reliability evaluation improve robustness against unreliable task responses while preserving benign task utility?

4.1 Experimental Setup

Models. Following KABB (Zhang et al., 2025b), we construct 20 agent profiles across 12 knowledge domains and distribute them over N federated clients. Each client hosts a local agent with private data, an experience buffer, and personalized privacy rules, while multiple clients may share the same profile. Agent profiles use domain specific prompts with one of seven LLM backbones: MiniMax2.5 (Chen et al., 2026), DeepSeek-V3 (Liu et al., 2024), Qwen2-72B-Instruct (Bai et al., 2023), GPT-4o (Hurst et al., 2024), DeepSeek-R1 (Guo et al., 2025), LLaMA-3-70B-Instruct (Adams et al., 2024), or Kimi-K2.5 (Team et al., 2026).

Dataset and Evaluation Metrics. We evaluate PFedMA on mathematical reasoning benchmarks, including GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021), and AIME24 (Muennighoff et al., 2025); clinical QA benchmarks, including MedQA and MedMCQA (Tang et al., 2024); instruction following with AlpacaEval 2.0 (Zhang et al., 2025b); and privacy sensitive web agent tasks with SeeAct on Mind2Web-SC (Xiang et al., 2025). We report ASR, RR, PNA, and NRP following ASB (Zhang et al., 2024), together with BP to evaluate benign task utility with defense mechanisms enabled, together with LPA, LPP, LPR, EA, FRA, and LR for privacy evaluation (Xiang et al., 2025; Shao et al., 2024). Metric definitions are provided in Appendix H.

Baselines. We conduct a comparative analysis of the PFedMA algorithm against other MAS methods. For the single agent baseline, the primary goal is to demonstrate that our method's performance is not due to the server LLM model, but rather to the effectiveness of the federated MAS. For federated MAS, we compare with Fed-ICL (Wang et al., 2025) and EPEAgents (Shi et al., 2025). For centralized MAS, we compare with Router-R1 (Zhang et al., 2025a), KABB (Zhang et al., 2025b), and OPTAGENT (Bi et al., 2025). All multi-agent methods use the same pool of 20 experts and the same number of iteration rounds to isolate the effect of orchestration and aggregation.

4.2 Main Results and Analysis

Table 1: Performance and privacy comparison of different frameworks. Best results are shown in **bold**, and second best results are underlined. For Avg calculation, lower is better metrics are converted using $100 - score$ before averaging.

RQ1: Performance comparison across multiple benchmarks							
Method	Math	GSM8K	AIME24	MedQA	AlpacaEval 2.0	MedMCQA	Avg
Single Agent	$42.63_{\pm 3.51}$	$79.11_{\pm 1.39}$	$6.67_{\pm 0.73}$	$63.35_{\pm 4.31}$	$58.51_{\pm 0.18}$	$57.82_{\pm 0.21}$	$51.35_{\pm 1.72}$
Router-R1	$76.06_{\pm 2.73}$	$84.02_{\pm 2.91}$	$73.40_{\pm 4.34}$	$75.61_{\pm 1.71}$	$75.43_{\pm 3.94}$	$69.57_{\pm 1.07}$	$75.68_{\pm 2.78}$
KABB	$77.51_{\pm 1.12}$	$85.66_{\pm 0.60}$	$71.93_{\pm 3.00}$	$\mathbf{78.15_{\pm 0.73}}$	$72.38_{\pm 4.30}$	$\mathbf{72.05_{\pm 1.70}}$	$76.28_{\pm 1.91}$
OPTAGENT	$75.02_{\pm 1.27}$	$86.26_{\pm 4.31}$	$62.43_{\pm 2.43}$	$66.27_{\pm 1.21}$	$69.01_{\pm 1.78}$	$64.82_{\pm 0.91}$	$70.64_{\pm 1.99}$
EPEAgents	$70.03_{\pm 0.39}$	$80.10_{\pm 1.76}$	$68.57_{\pm 3.16}$	$77.20_{\pm 1.99}$	$70.88_{\pm 1.36}$	$63.54_{\pm 0.81}$	$71.72_{\pm 1.58}$
Fed-ICL	$79.33_{\pm 0.83}$	$85.04_{\pm 1.04}$	$74.78_{\pm 1.51}$	$76.03_{\pm 3.00}$	$73.51_{\pm 0.29}$	$69.34_{\pm 0.93}$	$76.34_{\pm 1.27}$
PFedMA	$\mathbf{84.53_{\pm 2.51}}$	$\mathbf{86.60_{\pm 0.81}}$	$\mathbf{76.67_{\pm 1.83}}$	$77.91_{\pm 0.84}$	$\mathbf{77.65_{\pm 0.18}}$	$69.80_{\pm 0.53}$	$\mathbf{78.86_{\pm 1.12}}$

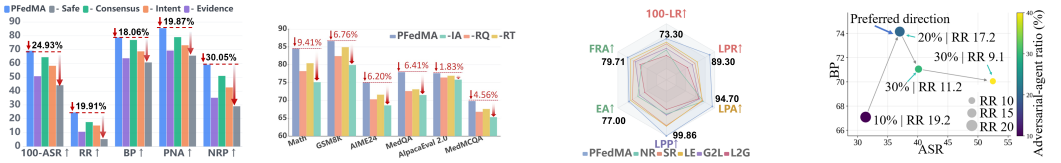
RQ2: Evaluation of privacy processed and performance							
Method	LR ↓	LPR ↑	LPA ↑	LPP ↑	EA ↑	FRA ↑	Avg ↑
Single Agent	$43.40_{\pm 0.62}$	$69.00_{\pm 1.14}$	$83.50_{\pm 0.87}$	$98.60_{\pm 0.35}$	$68.00_{\pm 1.26}$	$62.50_{\pm 0.94}$	$73.03_{\pm 0.86}$
Router-R1	$39.08_{\pm 0.78}$	$73.73_{\pm 1.31}$	$85.70_{\pm 0.69}$	$96.90_{\pm 0.52}$	$73.70_{\pm 1.08}$	$73.05_{\pm 0.83}$	$77.33_{\pm 0.87}$
KABB	$35.20_{\pm 0.91}$	$81.64_{\pm 0.74}$	$90.00_{\pm 0.58}$	$98.10_{\pm 0.41}$	$76.50_{\pm 1.17}$	$69.50_{\pm 0.96}$	$80.09_{\pm 0.80}$
OPTAGENT	$37.30_{\pm 0.85}$	$80.04_{\pm 1.09}$	$90.00_{\pm 0.63}$	$95.32_{\pm 0.22}$	$\mathbf{80.00_{\pm 0.48}}$	$72.10_{\pm 0.77}$	$80.03_{\pm 0.67}$
EPEAgents	$33.30_{\pm 0.66}$	$86.50_{\pm 0.82}$	$93.37_{\pm 0.54}$	$99.01_{\pm 0.19}$	$70.50_{\pm 1.21}$	$75.63_{\pm 0.73}$	$81.95_{\pm 0.69}$
Fed-ICL	$32.60_{\pm 0.71}$	$84.00_{\pm 0.95}$	$92.00_{\pm 0.61}$	$97.24_{\pm 0.17}$	$75.60_{\pm 1.04}$	$73.30_{\pm 0.88}$	$81.59_{\pm 0.73}$
PFedMA	$\mathbf{26.70_{\pm 0.43}}$	$\mathbf{89.30_{\pm 0.69}}$	$\mathbf{94.70_{\pm 0.38}}$	$\mathbf{99.86_{\pm 0.12}}$	$77.00_{\pm 0.97}$	$\mathbf{79.71_{\pm 0.55}}$	$\mathbf{85.65_{\pm 0.52}}$

Table 2: Robustness of different frameworks under adversarial attacks

RQ3: Evaluation of system security and robustness						
Method	ASR ↓	RR ↑	BP ↑	PNA ↑	NRP ↑	Avg ↑
Single Agent	$54.82_{\pm 1.23}$	$5.34_{\pm 2.41}$	$77.75_{\pm 3.84}$	$66.57_{\pm 2.03}$	$30.08_{\pm 3.20}$	$44.98_{\pm 2.54}$
Router-R1	$52.07_{\pm 0.81}$	$9.10_{\pm 0.36}$	$76.73_{\pm 0.74}$	$76.70_{\pm 0.65}$	$36.76_{\pm 0.88}$	$49.44_{\pm 0.69}$
KABB	$36.20_{\pm 0.69}$	$18.50_{\pm 0.54}$	$75.29_{\pm 0.83}$	$\mathbf{88.00_{\pm 0.47}}$	$56.14_{\pm 0.76}$	$60.35_{\pm 0.66}$
OPTAGENT	$50.07_{\pm 0.77}$	$11.50_{\pm 2.49}$	$69.40_{\pm 0.91}$	$73.54_{\pm 0.58}$	$36.72_{\pm 1.02}$	$48.22_{\pm 1.15}$
EPEAgents	$32.01_{\pm 0.52}$	$22.13_{\pm 0.61}$	$76.57_{\pm 1.72}$	$82.84_{\pm 0.55}$	$56.32_{\pm 0.69}$	$61.17_{\pm 0.82}$
Fed-ICL	$49.30_{\pm 0.84}$	$13.00_{\pm 0.43}$	$74.75_{\pm 3.79}$	$74.71_{\pm 1.66}$	$37.88_{\pm 1.73}$	$50.21_{\pm 1.69}$
PFedMA	$\mathbf{30.80_{\pm 0.45}}$	$\mathbf{24.60_{\pm 0.58}}$	$\mathbf{78.80_{\pm 0.63}}$	$85.50_{\pm 0.49}$	$\mathbf{59.17_{\pm 0.71}}$	$\mathbf{63.45_{\pm 0.57}}$

Cross Round Feedback and Iterative Orchestration (RQ1). Table 1 compares PFedMA with representative single-agent, centralized MAS, and federated MAS baselines under the same expert pool and collaboration-round budget. PFedMA achieves the highest average score of 78.86, outperforming the strongest baseline, Fed-ICL, by 2.52 points. It ranks first on MATH, GSM8K, AIME24, and

AlpacaEval 2.0, and second on MedQA and MedMCQA, indicating strong performance across the evaluated tasks. Figure 9a shows that accuracy generally increases across collaboration rounds, suggesting that useful reasoning information is progressively accumulated. The ablation results in Figure 2b further show that removing query optimization or either historical routing state consistently degrades performance. These results indicate that cross round state evolution, query adaptation, and learned orchestration each contribute to the overall improvement. Figure 6c provides complementary evidence that PFedMA adaptively selects agents according to task requirements: mathematical agents are selected more frequently for mathematical benchmarks, while medical agents are favored for medical tasks. Overall, these results support the effectiveness of feedback driven iterative orchestration in progressively refining the reasoning state and adapting subsequent collaboration decisions.



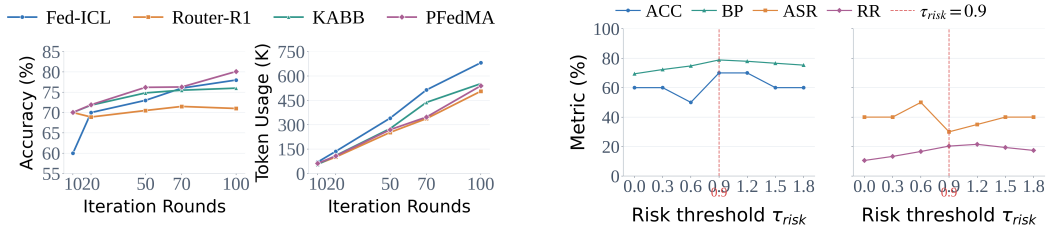
(a) Ablation of structured reliability evaluation

(b) Feedback Driven Orchestration Ablation

(c) Federated Rule Co-Evolution Ablation

(d) Adversarial Robustness Utility Trade off

Figure 2: Ablation studies of the reliability module, iterative orchestration loop, federated privacy rule co-evolution loop and robustness utility trade off across adversarial ratios.



(a) Accuracy and Token Usage by Round

(b) Sensitivity Analysis of the Risk Threshold τ

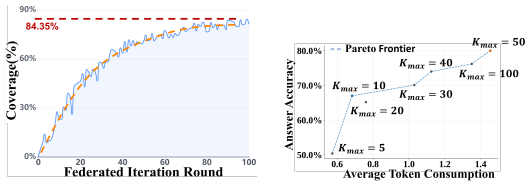
Figure 3: Performance Evaluation under Iterative Optimization and Risk Threshold Variation.

Cross Client Privacy Rule Coordination and Personalized Adaptation (RQ2). Table 1 shows that PFedMA achieves the lowest leakage rate of 26.70, which is 5.90 points lower than the second best result. It also obtains the best LPR, LPA, LPP, and FRA scores while achieving the second best EA, indicating that the improved privacy protection does not come at the cost of substantially degraded task utility. The lower leakage rate, together with the higher FRA, indicates that PFedMA strengthens measured privacy rule enforcement while preserving task utility under the evaluated setting. Figure 4a further shows that privacy rule coverage increases progressively across collaboration rounds. This trend is consistent with the intended role of federated privacy rule co-evolution, through which reusable rules identified locally are consolidated globally and subsequently adapted to client specific requirements. Together, these results provide empirical support for cross client privacy rule coordination and personalized adaptation while maintaining a favorable privacy utility balance.

Reliability Evaluation and Adversarial Robustness (RQ3). Table 2 compares PFedMA with representative baselines under prompt injection attacks. PFedMA achieves the lowest ASR of 30.80, reducing the attack success rate by 1.21 points compared with the second best method, EPEAgents. Meanwhile, PFedMA obtains the highest RR, BP, and NRP, and the second highest PNA, resulting in the best overall score of 63.45. These results indicate that the robustness improvement is not achieved by simply rejecting more responses. PFedMA remains effective on benign inputs while reducing the influence of adversarial responses. Overall, the results demonstrate a favorable robustness utility trade off under adversarial federated collaboration.

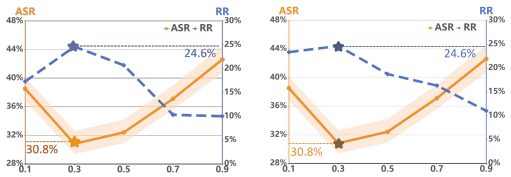
4.3 Ablation Study

Effectiveness of Feedback Driven Iterative Orchestration. Figure 2b evaluates the contributions of query optimization and the historical routing states. Specifically, the -RQ variant removes feed-



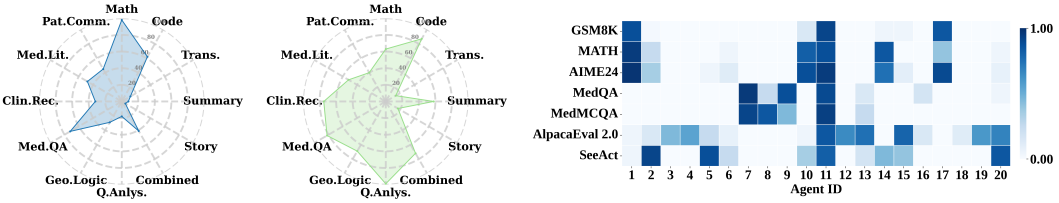
(a) Privacy Coverage Rate (b) Accuracy Efficiency

Figure 4: (a) Privacy rule coverage across rounds and (b) accuracy efficiency trade off across routing budgets.



(a) Hyperparameter λ_{pool} (b) Consensus α_{con}

Figure 5: Hyperparameters λ_{pool} and α_{con}. The x-axis represents the parameter values, and the y-axis shows the corresponding ASR and RR.



(a) Abilities of Agent 1 (b) Abilities of Agent 6 (c) Agent Selection Heatmap

Figure 6: Agent Abilities and Agent Selection. Full results are given in Appendix J.2.

back guided query optimization, the -IA variant removes the historical capability state C_i, and the -RT variant removes the historical reliability state S_i. Each ablation breaks a distinct pathway through which outcomes from the current round influence subsequent collaboration decisions. The consistent degradation of these variants demonstrates that PFedMA benefits from maintaining a closed cross round feedback loop, rather than merely repeating independent multi-agent interactions.

Necessity of Federated Privacy Rule Co-Evolution. Figure 2c compares PFedMA with five variants, where NR removes privacy rules, SR keeps rules static, LE updates rules only locally, G2L retains only global to local rule adaptation, and L2G retains only local to global rule aggregation. PFedMA achieves the privacy utility trade off, confirming the importance of combining personalized local adaptation with cross client rule synchronization. Figure 9c further shows how privacy rule effectiveness evolves across federated communication rounds. PFedMA consistently achieves the highest FRA, outperforming local only evolution, static rules, and the variant without privacy rules. Its persistent advantage across rounds indicates that local adaptation alone is insufficient to fully exploit reusable privacy knowledge. Together, these results demonstrate that bidirectional local to global rule consolidation and global to local personalized adaptation are both important for improving privacy rule effectiveness while preserving task utility.

Effectiveness of Multi Factor Reliability Evaluation. Figure 2a compares PFedMA with variants that remove the complete reliability evaluator or individual evaluation dimensions, including intent, evidence, and consensus. PFedMA achieves the best robustness utility profile, confirming the importance of reliability aware response assessment before aggregation. Removing the complete evaluator causes the largest overall degradation, while removing any individual dimension also weakens performance. These results indicate that relying on a single reliability signal is insufficient to consistently identify unreliable responses.

5 Conclusion and Limitations

We propose PFedMA, a federated multi-agent framework integrating feedback-driven orchestration with federated privacy rule co-evolution. The orchestration loop adapts agent selection based on evolving task requirements and historical reliability, while the privacy loop enables bidirectional rule evolution while preserving personalized constraints. Experiments across reasoning, privacy-sensitive, and adversarial tasks demonstrate improvements in task performance, privacy rule effectiveness, and robustness. PFedMA focuses on semantic privacy rule coordination under federated data isolation without providing formal privacy guarantees.

6 AI USE STATEMENT

In this study, we used generative artificial intelligence (AI) tools to assist with language editing and grammar correction and to improve the clarity, flow, and readability of the paper. During project development, we also used generative AI tools for limited coding assistance, debugging, and refinement of implementation details. Generative AI tools were not used to generate experimental results, fabricate data, conduct core scientific analyses, or make final judgments regarding the research methods, interpretation of results, or conclusions. Other disclosure items unrelated to these uses are not applicable to this study. All AI-assisted content was carefully reviewed, revised, and verified by the authors. The core research ideas, technical contributions, experimental design, and scientific claims presented in this paper were independently developed and established by the authors. All final decisions concerning the study design, implementation, experiments, analyses, and presentation of results were made by the authors. We take full responsibility for the final content of this work.

7 ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. We did not recruit human participants, collect new personal data, or conduct experiments involving human or animal subjects. Our experiments use existing benchmarks, including GSM8K, MATH, AIME24, MedQA, MedMCQA, AlpacaEval 2.0, and Mind2Web-SC. PFedMA aims to limit the disclosure of client-local information through adaptive privacy rules, while keeping raw local responses and complete personalized privacy rules on clients. We evaluate sensitive information leakage from generated outputs and assess adversarial robustness using benchmark-based prompt-injection attacks. These evaluations characterize empirical privacy and security performance and do not establish formal guarantees of differential privacy or cryptographic confidentiality.

8 REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide the implementation of PFedMA in an anonymous repository: `https://anonymous.4open.science/r/paper_PFedMA-2F44/`. The main paper and Appendices A–L detail the framework, agent configurations, response reliability evaluation, feedback-driven routing, privacy rule co-evolution, training settings, prompt templates, and evaluation protocols, supporting replication of our baseline comparisons, ablation studies, and analyses of privacy, adversarial robustness, client heterogeneity, efficiency, and hyperparameter sensitivity.

REFERENCES

Lisa C Adams, Daniel Truhn, Felix Busch, Felix Dorfner, Jawed Nawabi, Marcus R Makowski, and Keno K Bressem. Llama 3 challenges proprietary state-of-the-art large language models in radiology board–style examination questions. *Radiology*, 312(2):e241191, 2024.

Diogo Almeida. Introducing system one models & jev. TypeSafe AI, 2026. URL `https://typesafe.ai/blog/introducing-system-one-models-and-jev`.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Zhenyu Bi, Meng Lu, Yang Li, Swastik Roy, Weijie Guan, Morteza Ziyadi, and Xuan Wang. Optagent: Optimizing multi-agent llm interactions through verbal reinforcement learning for enhanced reasoning. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 1713–1728, 2025.

Aili Chen, Aonian Li, Baichuan Zhou, Bangwei Gong, Binyang Jiang, Boji Dan, Changhao Zhang, Changqing Yu, Chao Wang, Cheng Ma, et al. The minimax-m2 series: Mini activations unleashing max real-world intelligence. *arXiv preprint arXiv:2605.26494*, 2026.

Minghui Chen, Ruinan Jin, Wenlong Deng, Yuanyuan Chen, Zhi Huang, Han Yu, and Xiaoxiao Li. Can textual gradient work in federated learning? *arXiv preprint arXiv:2502.19980*, 2025.

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 *The Twelfth International Conference on Learning Representations*, 2023.

Hyeong Kyu Choi, Xiaojin Zhu, and Sharon Li. Debate or vote: Which yields better decisions in multi-agent large language models? *arXiv preprint arXiv:2508.17536*, 2025.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Wentao Deng, Jiahuan Pei, Zhiwei Xu, Zhaochun Ren, Zhumin Chen, and Pengjie Ren. Belief-calibrated multi-agent consensus seeking for complex nlp tasks. *arXiv preprint arXiv:2510.06307*, 2025.

Haonan Duan, Adam Dziedzic, Mohammad Yaghini, Nicolas Papernot, and Franziska Boenisch. On the privacy risk of in-context learning. *arXiv preprint arXiv:2411.10512*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*, 2024.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Tianyu Hu, Zhen Tan, Song Wang, Huaizhi Qu, and Tianlong Chen. Multi-agent debate for llm judges with adaptive stability detection. *arXiv preprint arXiv:2510.12697*, 2025.

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.

Anita Keshmirian, Razan Baltaji, Babak Hemmatian, Hadi Asghari, and Lav R Varshney. Many llms are more utilitarian than one. *arXiv preprint arXiv:2507.00814*, 2025.

Magdalena Kossek and Margareta Stefanovic. Survey of recent results in privacy-preserving mechanisms for multi-agent systems. *Journal of Intelligent & Robotic Systems*, 110(3):129, 2024.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori B Hashimoto. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20286–20332, 2025.

Anudit Nagar, Cuong Tran, and Ferdinando Fioretto. A privacy-preserving and trustable multi-agent learning framework. *arXiv preprint arXiv:2106.01242*, 2021.

Ashwinee Panda, Tong Wu, Jiachen Wang, and Prateek Mittal. Differentially private in-context learning. In *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023.

Chen Qiu, Xingyu Li, Chaithanya Kumar Mummadi, Madan Ravi Ganesh, Zhenzhen Li, Lu Peng, and Wan-Yi Lin. Text-driven prompt generation for vision-language models in federated learning. *arXiv preprint arXiv:2310.06123*, 2023.

Yijia Shao, Tianshi Li, Weiyan Shi, Yanchen Liu, and Diyi Yang. Privacylens: Evaluating privacy norm awareness of language models in action. *Advances in Neural Information Processing Systems*, 37:89373–89407, 2024.

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.

Yiliu Sun, Zicheng Zhao, Sheng Wan, and Chen Gong. Cortexdebate: Debating sparsely and equally for multi-agent debate. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 9503–9523, 2025.

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. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 599–621, 2024.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, et al. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

Nguyen Truong, Kai Sun, Siyao Wang, Florian Guitton, and YiKe Guo. Privacy preservation in federated learning: An insightful survey from the gdpr perspective. *Computers & Security*, 110: 102402, 2021.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692*, 2024.

Ruhan Wang, Zhiyong Wang, Chengkai Huang, Rui Wang, Tong Yu, Lina Yao, John Lui, and Dongruo Zhou. Federated in-context learning: Iterative refinement for improved answer quality. *arXiv preprint arXiv:2506.07440*, 2025.

Yaqi Wang, Jianquan Lu, Wei Xing Zheng, and Kaibo Shi. Privacy-preserving consensus for multi-agent systems via node decomposition strategy. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 68(8):3474–3484, 2021.

Panlong Wu, Kangshuo Li, Junbao Nan, and Fangxin Wang. Federated in-context llm agent learning. *arXiv preprint arXiv:2412.08054*, 2024.

Zhen Xiang, Linzhi Zheng, Yanjie Li, Junyuan Hong, Qinbin Li, Han Xie, Jiawei Zhang, Zidi Xiong, Chulin Xie, Nathaniel D Bastian, et al. Guardagent: safeguard llm agents via knowledge-enabled reasoning. In *ICML 2025 workshop on computer use agents*, 2025.

Mengxi Xiao, Ben Liu, He Li, Jimin Huang, Qianqian Xie, Xiaofen Zong, Mang Ye, and Min Peng. Moodangels: A retrieval-augmented multi-agent framework for psychiatry diagnosis. *arXiv preprint arXiv:2506.03750*, 2025.

Yanwei Yue, Guibin Zhang, Boyang Liu, Guancheng Wan, Kun Wang, Dawei Cheng, and Yiyan Qi. Masrouter: Learning to route llms for multi-agent systems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15549–15572, 2025.

Hanrong Zhang, Jingyuan Huang, Kai Mei, Yifei Yao, Zhenting Wang, Chenlu Zhan, Hongwei Wang, and Yongfeng Zhang. Agent security bench (asb): Formalizing and benchmarking attacks and defenses in llm-based agents. *arXiv preprint arXiv:2410.02644*, 2024.

Haozhen Zhang, Tao Feng, and Jiaxuan You. Router-r1: Teaching llms multi-round routing and aggregation via reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a.

Jusheng Zhang, Zimeng Huang, Yijia Fan, Ningyuan Liu, Mingyan Li, Zhuojie Yang, Jiawei Yao, Jian Wang, and Keze Wang. Kabb: Knowledge-aware bayesian bandits for dynamic expert coordination in multi-agent systems. *arXiv preprint arXiv:2502.07350*, 2025b.

Yanzhe Zhang and Diyi Yang. Searching for privacy risks in llm agents via simulation. In *International Conference on Learning Representations*, volume 2026, pp. 98142–98176, 2026.

A ALGORITHM

In this section, we mainly introduce the core algorithm in our paper.

Algorithm 1 Execution Workflow of PFedMA

Input: initial task q^1, clients set $\mathcal{A} = \{A_1, \ldots, A_N\}$, domain set $\mathcal{D}omain = \{d_1, \ldots, d_L\}$, domain prototypes $\{\rho_d\}_{d \in \mathcal{D}omain}$, maximum number of rounds $T_{\max}$, maximum number of clients per round $K_{\max}$, risk threshold τ_{risk}, consensus weight α_{con}, global privacy rule repository Ω^1_{global}, and personalized privacy rule repositories $\{\Omega^1_{i,\text{local}}\}_{i=1}^N$
Initialize $t \leftarrow 1$ $C_i^1 \leftarrow \mathbf{0}$, $S_i^1 \leftarrow \mathbf{0}$, $z^0 \leftarrow \emptyset$, $n_i^1 \leftarrow 0$ for all $A_i \in \mathcal{A}$
Each agent A_i maintains local experience $\mathcal{X}_i$ and routing state $\mathbf{M}_i^1 = [\mathbf{C}_i^1 \| \mathbf{S}_i^1]$
while $t \leq T_{\max}$ **do**
  Apply the global privacy rule repository to the current query: $\tilde{q}^t \leftarrow P_{\text{in}}(q^t, \Omega^t_{\text{global}})$
  Identify the most relevant domain: $d^t_\star \leftarrow \arg\max_{d \in \mathcal{D}omain} \text{Cos}(\tilde{q}^t, \rho_d)$
  Retrieve the corresponding candidate agent set $\mathcal{A}_{d^t_\star}$
  Select participating agents: $\mathcal{E}^t \leftarrow \pi_\theta\left(\mathcal{A}_{d^t_\star} \mid \tilde{q}^t, z^{t-1}, \{\mathbf{M}_i^t\}_{A_i \in \mathcal{A}_{d^t_\star}}\right), 1 \leq |\mathcal{E}^t| \leq K_{\max}$
  if π_θ selects the termination action **then**
    break
  end if
  for all $A_i \in \mathcal{E}^t$ **do**
    Generate the local response: $r_i^t \leftarrow A_i(\tilde{q}^t, D_i, \mathcal{X}_i)$
    Apply personalized privacy processing: $\tilde{r}_i^t \leftarrow P_{\text{resp},i}(r_i^t, \Omega^t_{i,\text{local}})$
  end for
  for all $A_i \in \mathcal{E}^t$ **do**
    $\text{Int}_i^t \leftarrow \text{IntentAnomaly}(\tilde{r}_i^t)$
    $\text{Evd}_i^t \leftarrow \text{EvidenceAnomaly}(\tilde{r}_i^t)$
    $\text{Con}_i^t \leftarrow \text{ConsensusAnomaly}(\tilde{r}_i^t)$
    $\text{Risk}_i^t \leftarrow \text{Int}_i^t + \text{Evd}_i^t + \alpha_{\text{con}}\text{Con}_i^t$
  end for
  Retain responses satisfying the reliability threshold: $\hat{\mathcal{E}}^t \leftarrow \left\{A_i \in \mathcal{E}^t \mid \text{Risk}_i^t \leq \tau_{\text{risk}}\right\}$
  if $\hat{\mathcal{E}}^t \neq \emptyset$ **then**
    Update the global reasoning state: $z^t \leftarrow \text{Fusion}\left(\{\tilde{r}_i^t\}_{A_i \in \hat{\mathcal{E}}^t}, z^{t-1}\right)$
  else
    $z^t \leftarrow z^{t-1}$
  end if
  Update the routing states $\{\mathbf{M}_i^{t+1} = \left[\mathbf{C}_i^{t+1} \| S_i^{t+1}\right]\}_{A_i \in \mathcal{E}^t}$ using the observed capability and reliability feedback
  Construct generalized local rule candidates from newly identified personalized privacy constraints
  Build the candidate pool $\mathcal{B}^t$ and obtain the supported global candidate set $\mathcal{G}^t$
  Collect rule-level feedback F^t and update the global repository: $\Omega^{t+1}_{\text{global}} \leftarrow \text{RuleUpdate}\left(\Omega^t_{\text{global}}, \mathcal{G}^t, F^t\right)$
  Redistribute updated global rules and selectively update each personalized repository $\Omega^{t+1}_{i,\text{local}}$
  Refine the query for the next round: $q^{t+1} \leftarrow \text{Refine}(q^t, z^t)$
  $t \leftarrow t + 1$
end while
Output: Final answer derived from the latest global reasoning state

B COMPARISON WITH RELATED FRAMEWORKS

As summarized in Table 3, existing methods primarily optimize expert selection, sequential routing, or federated response refinement in isolation. In contrast, PFedMA establishes a closed loop

Table 3: Comparison of PFedMA with representative multi-agent and federated multi-agent frameworks. PFedMA differs from prior methods by jointly coupling iterative reasoning, agent reliability, and heterogeneous privacy constraints within a closed loop orchestration process.

Method	Setting	Expert Coordination	Iterative Refinement	Reliability Feedback	Privacy Personalization	Privacy Rules Co-evolution
Router-R1	Centralized MAS	RL-based dynamic routing	✓	–	–	–
KABB	Centralized MAS	Knowledge aware expert selection	Partial	–	–	–
Fed-ICL	Federated MAS	Federated response aggregation	✓	–	–	–
PFedMA	**Federated MAS**	**Reliability aware dynamic routing**	✓	✓	✓	✓

orchestration process in which reasoning progress, response reliability, and heterogeneous privacy constraints jointly influence subsequent collaboration decisions.

C Structured Response Evaluation Details

C.1 Structured Response Format and Example

The agent's response follows a structured format consisting of three components: final_answer, claim, and evidence. At round t, each selected client A_i returns a privacy rule based information minimization response using the following structured format:

```
{
  "final_answer": "...",
  "claims": [
    {
      "claim": "...",
      "evidence": ["...", "..."]
    },{
      "claim": "...",
      "evidence": ["...", "..."]
    }
  ],
  "domain_fit": 0.0,
  "self_confidence": 0.0
}
```

The corresponding mathematical representation is

$$\tilde{r}_i^t = \left\langle a_i^t, \Gamma_i^t, \mathcal{L}_i^t \right\rangle, \tag{12}$$

The structured response separates the final conclusion from the claim level information used for subsequent reliability evaluation. Here, a_i^t denotes the final conclusion returned by agent A_i at round t.

$$\Gamma_i^t = \gamma_{i,n}^t{}_{n=1}^{|\Gamma_i^t|}, \tag{13}$$

where $\gamma_{i,n}^t$ denotes the n-th explicit claim in the response, and $|\Gamma_i^t|$ denotes the number of claims.

Each claim is paired with a corresponding evidence set, such that

$$\mathcal{L}_i^t = L_{i,n}^t{}_{n=1}^{|\Gamma_i^t|}, \tag{14}$$

where $L_{i,n}^t$ contains the evidence provided specifically for claim $\gamma_{i,n}^t$. In particular,

$$L_{i,n}^t = \left\{ \epsilon_{i,n,k}^t \right\}_{k=1}^{m_{i,n}^t}, \tag{15}$$

where $\epsilon_{i,n,k}^t$ denotes the k-th evidence item associated with the n-th claim, and $m_{i,n}^t$ denotes the number of evidence items provided for that claim. Thus, the structured representation establishes an explicit claim evidence correspondence: each $\gamma_{i,n}^t$ is evaluated together with its own evidence set $L_{i,n}^t$, while different claims may be supported by different numbers of evidence items.

The fields `domain_fit` and `self_confidence` are self reported auxiliary metadata in $[0, 1]$. Because these values may be manipulated by unreliable clients, they are not directly treated as reliability scores.

Table 4: Implementation and training status of the response evaluators.

Component	Implementation	Shares π_θ	Trainable
Intent evaluator	Frozen LLM judge	No	No
Evidence evaluator	Frozen LLM judge	No	No
Consensus evaluator	Semantic similarity	No	No
Orchestrator	RL fine tuned LLM	–	Yes

Example. **[AGE]** year old **[GENDER]** with transitional cell carcinoma of the bladder comes to the physician because of a **[TIME]** day history of ringing sensation in his ear. He received this first course of neoadjuvant chemotherapy **[TIME]**. Pure tone audiometry shows a sensorineural hearing loss of **[NUMBER]** dB. The expected beneficial effect of the drug that caused this patient's symptoms is most likely due to which of the following actions? An example of a client response is shown in the Figure 7.

Final Answer Format.

Final Answer

Cross-linking of DNA.

Claims

Claim 1. The patient is taking a chemotherapeutic agent used to treat bladder cancer, most likely a platinum-based drug like Cisplatin.

Evidence:

- transitional cell carcinoma of the bladder
- first course of neoadjuvant chemotherapy

Claim 2. The mechanism of action, or beneficial effect, of Cisplatin is the formation of intra- and inter-strand DNA cross-links.

Evidence:

- expected beneficial effect of the drug that caused this patient's symptoms

Figure 7: Format of the client's response.

C.2 Evaluator Implementation and Parameter Isolation

The response evaluators are separated from the trainable orchestrator π_θ(Qwen2.5-3B-Instruct). Specifically, the intent and evidence anomaly scores are produced by a frozen judge model using fixed evaluation prompts and predefined scoring criteria. The judge does not share trainable parameters with π_θ, and its parameters remain unchanged during both reinforcement learning training and inference. The intent and evidence evaluators use the same frozen judge model (Qwen2.5-3B-Instruct) backbone but different task specific instructions.

The consensus anomaly does not invoke the judge model. It is computed deterministically from the semantic similarity between the conclusions of the participating agents. Therefore, the three anomaly dimensions do not introduce additional trainable components into PFedMA.

The judge receives only the response fields required by the corresponding evaluator. The intent evaluator receives the response sentences and the predefined intent risk categories, whereas the evidence evaluator receives each claim and its associated evidence. It does not observe the ground truth answer, the routing reward, or the orchestrator's subsequent decision. The resulting anomaly scores are used by the server for response filtering and agent state updates, but no evaluator gradient is propagated to the orchestrator.

C.3 Anomaly Definitions and Boundary Cases

Pooling Operator For any non empty set of scores $\{x_j\}_{j=1}^{J}$, we define

$$\mathrm{Pool}_{\lambda_{pool}}\left(\{x_j\}_{j=1}^{J}\right) = \lambda_{pool} \max_{1\le j\le J} x_j + (1-\lambda_{pool})\frac{1}{J}\sum_{j=1}^{J} x_j, \tag{16}$$

where J denotes the number of fine-grained anomaly scores being aggregated, x_j denotes the j-th anomaly score, and $\lambda_{\mathrm{pool}} \in [0,1]$ controls the trade-off between the most severe anomaly and the average anomaly level.

Intent Anomaly. Let $\Xi_i^t = \left\{\xi_{i,j}^t\right\}_{j=1}^{J_i^t}$ denote the sentence level units extracted from $\tilde{r}_i^t$. The predefined intent risk set is

$$\nu_{\mathrm{int}} = \{\text{instruction overriding}, \text{false authority}, \text{aggregator manipulation}, \text{goal substitution}\}. \tag{17}$$

For each response unit, the evaluator produces an intent risk score

$$u_{i,j}^t = \mathrm{Eval}_{\mathrm{int}}\left(\xi_{i,j}^t, \nu_{\mathrm{int}}\right) \in [0,1]. \tag{18}$$

The response level intent anomaly is

$$Int_i^t = \mathrm{Pool}_{\lambda_{\mathrm{pool}}}\left(\{u_{i,j}^t\}_{j=1}^{J_i^t}\right). \tag{19}$$

Evidence Anomaly. Privacy rule based information minimization adjustment or abstracts sensitive details while aiming to preserve the evidence supporting each claim. The evaluator assesses the support provided by the privacy processed evidence; omitted sensitive details are not themselves treated as anomalies when the remaining information sufficiently supports the claim. Verifiability refers to the judge's assessment based on the supplied text, rather than independent verification of external sources. For claim $\gamma_{i,n}^t$, let

$$\mathcal{L}_{i,n}^t = \{\varepsilon_{i,n,k}^t\}_{k=1}^{m_{i,n}^t}, \tag{20}$$

where $\varepsilon_{i,n,k}^t$ denotes the k-th evidence item associated with the n-th claim, and $m_{i,n}^t$ denotes the number of evidence items provided for that claim. For $m_{i,n}^t > 0$, let $\mathrm{Sup}_{i,n,k}^t \in [0,1]$ and $\mathrm{Ver}_{i,n,k}^t \in [0,1]$ denote the support strength and verifiability of evidence item $\varepsilon_{i,n,k}^t$, respectively. We define

$$\overline{\mathrm{Sup}}_{i,n}^t = \frac{1}{m_{i,n}^t}\sum_{k=1}^{m_{i,n}^t} \mathrm{Sup}_{i,n,k}^t, \tag{21}$$

$$\mathrm{Ver}_{i,n}^t = \max_{1\le k\le m_{i,n}^t} \mathrm{Ver}_{i,n,k}^t. \tag{22}$$

The internal coherence of the evidence set is

$$\mathrm{Coh}_{i,n}^t = \begin{cases} \dfrac{2}{m_{i,n}^t\left(m_{i,n}^t-1\right)} \displaystyle\sum_{1\le k_1<k_2\le m_{i,n}^t} \mathrm{Cons}\left(\varepsilon_{i,n,k_1}^t, \varepsilon_{i,n,k_2}^t\right), & m_{i,n}^t > 1, \\ 1, & m_{i,n}^t = 1, \\ 0, & m_{i,n}^t = 0. \end{cases} \tag{23}$$

The claim level evidence quality is

$$\mathrm{Qual}_{i,n}^{\mathrm{evd}\ t} = \begin{cases} \dfrac{\overline{\mathrm{Sup}}_{i,n}^t + \mathrm{Ver}_{i,n}^t + \mathrm{Coh}_{i,n}^t}{3}, & m_{i,n}^t > 0, \\ 0, & m_{i,n}^t = 0. \end{cases} \tag{24}$$

Accordingly, the claim level evidence anomaly is

$$\mathrm{Anom}_{i,n}^{\mathrm{evd}\ t} = 1 - \mathrm{Qual}_{i,n}^{\mathrm{evd}\ t}. \tag{25}$$

The response level evidence anomaly is

$$\mathrm{Evd}_i^t = \begin{cases} \mathrm{Pool}_{\lambda_{\mathrm{pool}}}\left(\left\{\mathrm{Anom}_{i,n}^{\mathrm{evd},t}\right\}_{n=1}^{|\Gamma_i^t|}\right), & |\Gamma_i^t| > 0, \\ 1, & |\Gamma_i^t| = 0. \end{cases} \tag{26}$$

When no explicit claim is provided, i.e., $|\Gamma_i^t| = 0$, the evidence anomaly is set to its maximum value of 1. This boundary condition prevents an empty claim set from bypassing evidence-quality evaluation and is consistent with the structured response format, in which claims and their associated evidence provide the basis for evidence reliability assessment.

Consensus Anomaly. The consensus anomaly measures whether the conclusion of client A_i is supported by the conclusions of the other selected clients:

$$Con_i^t = \begin{cases} 1 - \dfrac{1}{|\mathcal{E}^t| - 1} \displaystyle\sum_{\substack{A_{i'} \in \mathcal{E}^t \\ i' \neq i}} \mathrm{Sim}\left(a_i^t, a_{i'}^t\right), & |\mathcal{E}^t| > 1, \\ 0, & |\mathcal{E}^t| = 1. \end{cases} \tag{27}$$

where $\mathrm{Sim}(\cdot,\cdot) \in [0,1]$ denotes normalized semantic similarity. A larger value therefore indicates stronger deviation from peer conclusions. The singleton case is assigned zero because consensus cannot be assessed from only one response.

C.4 Overall Anomaly Score and Robust Aggregation

The three anomaly dimensions are combined into an overall response anomaly score:

$$\mathcal{R}isk_i^t = Int_i^t + Evd_i^t + \alpha_{\mathrm{con}} Con_i^t. \tag{28}$$

Responses whose anomaly scores exceed the tolerance threshold are excluded from aggregation:

$$\widehat{\mathcal{E}}^t = \left\{A_i \in \mathcal{E}^t \mid \mathcal{R}isk_i^t \leq \tau_{\mathrm{risk}}\right\}. \tag{29}$$

The orchestrator then aggregates the retained privacy rule based information minimization responses to update the accumulated collaboration state:

$$z^t = \begin{cases} \mathrm{Fusion}\left(\{\tilde{r}_i^t\}_{A_i \in \widehat{\mathcal{E}}^t}, z^{t-1}\right), & \widehat{\mathcal{E}}^t \neq \varnothing, \\ z^{t-1}, & \widehat{\mathcal{E}}^t = \varnothing. \end{cases} \tag{30}$$

D Agent Routing State Update

After each orchestration round, PFedMA updates the routing metadata of the participating agents based on their observed historical capability and historical response reliability.

For agent A_i with at least one previously retained response, define

$$\tau_i(t) = \max\{s < t \mid A_i \in \widehat{\mathcal{E}}^s\} \tag{31}$$

Thus, $\tau_i(t)$ indexes the most recent round before t in which the response of client A_i was retained by the reliability filter. Denote the most recent round before t in which the response of agent A_i was retained by the reliability filter. Here, s indexes previous collaboration rounds. This definition allows the historical capability update to compare the current response with the most recent reliable response of the same agent while naturally accommodating intermittent agent participation across rounds.

If the current response passes the reliability filter and the agent has a previously retained response, the orchestrator compares the current response with that agent's most recent retained response from an earlier round to obtain a response level improvement signal and an agent specific capability increment vector.

$$f_i^t = \pi_\theta^{\mathrm{fb}}\left(\tilde{r}_i^t \mid \tilde{r}_i^{\tau_i(t)}\right), \qquad \bar{\mathbf{v}}_i^t = \pi_\theta^{\mathrm{cap}}\left(\tilde{r}_i^t \mid \tilde{r}_i^{\tau_i(t)}\right). \tag{32}$$

Here, π_θ^{fb} and $\pi_\theta^{\mathrm{cap}}$ denote two comparative evaluations performed by the same orchestrator π_θ under different fixed evaluation instructions, rather than conditional likelihoods. The response level feedback signal $f_i^t \in \{0, 1\}$ indicates whether the current retained response improves upon the most recent retained response of the same client. The capability vector $\bar{\mathbf{v}}_i^t$ identifies the capability dimensions in which the current response improves upon that reference. Before collaboration begins, the historical capability state of each agent is initialized as a zero vector: $\mathbf{C}_i^1 = [0, \ldots, 0]^\top$.

For the first retained response of an agent, no previously retained response is available for comparison. This response therefore establishes the initial comparison reference, and no capability progress increment is assigned in that round.

For subsequent responses, the historical capability state is updated as

$$\mathbf{C}_i^{t+1} = \begin{cases} \mathbf{C}_i^t + f_i^t \bar{\mathbf{v}}_i^t, & A_i \in \widehat{\mathcal{E}}^t, \\ \mathbf{C}_i^t, & A_i \notin \widehat{\mathcal{E}}^t, \end{cases} \tag{33}$$

where $\widehat{\mathcal{E}}^t$ denotes the set of responses retained by the reliability filter. Therefore, responses excluded from aggregation do not receive positive progress credit.

The historical reliability state records the cumulative mean of the anomaly scores observed from the responses of agent A_i. Let n_i^t denote the number of responses from agent A_i that have been evaluated prior to round t. Before collaboration begins, the observation count and historical reliability state are initialized as

$$n_i^1 = 0, \qquad \mathbf{S}_i^1 = [0 \quad 0 \quad 0]^\top. \tag{34}$$

Whenever a new response from agent A_i is evaluated, the observation count is updated as

$$n_i^{t+1} = n_i^t + 1, \tag{35}$$

and the historical reliability state is updated as

$$\mathbf{S}_i^{t+1} = \frac{n_i^t \mathbf{S}_i^t + \left[\mathrm{Int}_i^t, \mathrm{Evd}_i^t, \mathrm{Con}_i^t\right]^\top}{n_i^t + 1}. \tag{36}$$

Here, Int_i^t, Evd_i^t, and Con_i^t denote the intent, evidence, and consensus anomaly scores of the current response, respectively. Unlike the historical capability update, every evaluated response contributes to the historical reliability state, including responses excluded from aggregation by the reliability filter. This allows unreliable responses to remain reflected in the agent's historical risk profile.

If agent A_i does not return a newly evaluated response at round t, its historical capability state, reliability states, and observation count remain unchanged. Finally, the updated routing state is defined as

$$M_i^{t+1} = \left[\mathbf{C}_i^{t+1} \,\|\, \mathbf{S}_i^{t+1}\right]. \tag{37}$$

Thus, the routing state jointly captures accumulated evidence of agent-specific historical capability and historical response reliability, allowing subsequent routing decisions to account for both observed collaboration progress and historical reliability.

E Reinforcement Learning Details

E.1 Reward Definitions

Let τ denote a collaboration trajectory containing T executed routing rounds, where $T_{\max}$ is the maximum routing round budget, $T \leq T_{\max}$, and $t \in \{1, \ldots, T\}$ indexes the executed rounds. As defined in Equation 6, the trajectory reward consists of a format reward, a terminal answer reward, and step-wise rewards that provide intermediate supervision for multi-round orchestration.

Format and final conclusion rewards. The format reward evaluates whether the orchestrator generated outputs satisfy the required structured format:

$$R_{\text{fmt}}(\tau) = \begin{cases} 0, & \text{if the required output format is satisfied,} \\ -1, & \text{otherwise.} \end{cases} \tag{38}$$

The terminal answer reward evaluates the correctness of the final answer using the corresponding benchmark evaluator:

$$R_{\text{ans}}(\tau) = \begin{cases} 1, & \text{if the terminal answer is judged correct,} \\ -1, & \text{otherwise.} \end{cases} \tag{39}$$

Cross-round reasoning progress. For a training instance with ground truth answer $y^\star$, let $\hat{y}^t$ denote the answer extracted from the aggregated reasoning state z^t using the corresponding benchmark's answer extraction procedure. Exact Match (EM) is defined as

$$\text{EM}_t = \mathbb{I}\big[\hat{y}^t = y^\star\big], \tag{40}$$

where $\mathbb{I}[\cdot]$ is the indicator function. To measure partial correctness, let $\mathcal{T}^t_{\text{pred}}$ and $\mathcal{T}_{\text{gt}}$ denote the token multisets obtained after applying the same benchmark specific normalization procedure to $\hat{y}^t$ and $y^\star$, respectively. Token level precision is

$$\text{Prec}_t = \begin{cases} \dfrac{\left|\mathcal{T}^t_{\text{pred}} \cap \mathcal{T}_{\text{gt}}\right|}{\left|\mathcal{T}^t_{\text{pred}}\right|}, & \left|\mathcal{T}^t_{\text{pred}}\right| > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{41}$$

Token level recall is

$$\text{Rec}_t = \begin{cases} \dfrac{\left|\mathcal{T}^t_{\text{pred}} \cap \mathcal{T}_{\text{gt}}\right|}{|\mathcal{T}_{\text{gt}}|}, & |\mathcal{T}_{\text{gt}}| > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{42}$$

Here, $\cap$ denotes multiset intersection, $|\cdot|$ denotes the number of tokens counted with multiplicity, precision is the proportion of predicted tokens that match the ground truth answer, and recall is the proportion of ground truth tokens covered by the prediction. The token level F1 score is

$$\text{F1}_t = \begin{cases} \dfrac{2\,\text{Prec}_t \text{Rec}_t}{\text{Prec}_t + \text{Rec}_t}, & \text{Prec}_t + \text{Rec}_t > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{43}$$

Since Prec_t and Rec_t are nonnegative, their sum is zero only when both scores are zero. The second case therefore avoids division by zero. Following (Zhang et al., 2025a), the intermediate answer quality score is

$$Q_t = \text{EM}_t + (1 - \text{EM}_t)\,\text{F1}_t. \tag{44}$$

Thus, $Q_t = 1$ for an exact match and $Q_t = \text{F1}_t$ otherwise. We set $Q_0 = 0$ and define the quality gain at round t as

$$\Delta Q_t = Q_t - Q_{t-1}. \tag{45}$$

Round level response risk. For each response $\tilde{r}^t_i$, we reuse the overall anomaly score $\mathcal{R}isk^t_i$. Since its three components are bounded in $[0, 1]$, its range is

$$B^t_{\text{risk}} = \begin{cases} 2, & |\mathcal{E}^t| = 1, \\ 2 + \alpha_{\text{con}}, & |\mathcal{E}^t| > 1. \end{cases} \tag{46}$$

Each response anomaly score is normalized by B^t_{risk}, and the normalized scores are averaged over the agents invoked at round t. The round-level response risk is

$$\overline{\mathcal{R}isk}_t = \frac{1}{|\mathcal{E}^t|} \sum_{A_i \in \mathcal{E}^t} \frac{\text{Risk}^t_i}{B^t_{\text{risk}}}. \tag{47}$$

where $\mathcal{E}^t$ is the set of agents invoked at round t. Here, $\alpha_{\text{con}} \geq 0$ is the consensus anomaly weight and $|\mathcal{E}^t|$ is the number of agents invoked at round t. Consequently, $\overline{\mathcal{R}isk}_t \in [0, 1]$ and a larger value indicates a less reliable set of responses.

Agent invocation cost. The normalized invocation cost is

$$\frac{|\mathcal{E}^t|}{K_{\max}}, \tag{48}$$

where $K_{\max}$ is the maximum number of agents that may be invoked in one routing round.

Combining the three components gives the step wise reward in Equation 7:

$$R^t_{\text{step}} = \Delta Q_t - \overline{\mathcal{R}isk}_t - \frac{|\mathcal{E}^t|}{K_{\max}}. \tag{49}$$

A routing step satisfies $1 \leq |\mathcal{E}^t| \leq K_{\max}$. If the orchestrator selects the termination action, no additional agent is invoked and no further step wise reward is generated.

The ground truth dependent quantities R_{ans}, Q_t, and ΔQ_t are used only during RL training. They are not available to the orchestrator during inference.

E.2 Policy Optimization

We optimize the server side orchestrator π_θ, where θ denotes its trainable parameters, using Proximal Policy Optimization (PPO) implemented with the veRL framework. At round t, the policy observes the privacy processed query $\tilde{q}^t$, the accumulated collaboration state z^{t-1}, and the capability and safety metadata M_i^t of the candidate agents. It then generates a structured routing action that either selects a subset $\mathcal{E}^t$ satisfying $1 \leq |\mathcal{E}^t| \leq K_{\max}$ or terminates the trajectory when no additional agent is required.

After the selected agents return Privacy rule based information minimization responses $\{\tilde{r}_i^t\}_{A_i \in \mathcal{E}^t}$, the server evaluates response reliability, aggregates the retained responses, and computes the training reward. PPO updates are performed centrally using the orchestrator visible trajectory. Raw client datasets $\mathcal{D}_i$, local experience buffers $\mathcal{X}_i$, complete personalized privacy rules $\Omega^t_{i,\text{local}}$, and unsanitized responses r_i^t remain client local and are not used in policy optimization. This training protocol preserves the federated data isolation boundary, but does not by itself provide a cryptographic or differential privacy guarantee.

E.3 Training Configuration

We train the Orchestrator π_θ using the veRL framework for reinforcement learning in LLMs, with Proximal Policy Optimization (PPO) as the default algorithm. Specifically, we use the training splits of MATH and MedQA for RL optimization of the Orchestrator, while all remaining benchmarks are evaluated in a zero shot manner without further RL fine tuning. During the RL training phase, the batch size is set to 64, with a maximum of 220 training steps. The anomaly evaluation parameters are set to $\lambda_{\text{pool}} = 0.3$ and $\alpha_{\text{con}} = 0.3$; their sensitivity is analyzed in Section J.5. Training this model requires 4 NVIDIA A800 GPUs, while inference requires 8 NVIDIA 3090 GPUs. Below are the hyperparameter settings for reinforcement learning in our work. We refer to (Zhang et al., 2025a) to set the hyperparameters required for the experiments.

Table 5: Hyperparameter Settings.

Hyperparameter	Value	Hyperparameter	Value
Learning Rate (Actor)	1e-5	Learning Rate (Critic)	1e-5
Total Batch Size	64	Mini-batch Size	64
Micro-batch Size	8	Max Training Steps	220
Max Routing Steps	100	Max Sequence Length	4096
Max Response Length	2048	Max Length for LLM API Response	1024
Tensor Parallel Size	1	GPU Utilization Ratio	0.8
Rollout Sampling Temperature (Train)	1.0	Rollout Sampling Temperature (Eval)	1.0

F FEDERATED PRIVACY RULE CO-EVOLUTION

Candidate Pool Construction. Only clients retained by the response reliability evaluation contribute new rule candidates. To prevent one client from dominating a semantic cluster, each client contributes at most one candidate to that cluster in a collaboration round. After this distinct client contribution check, explicit client identifiers are removed from the stored representations, and the candidates are randomly permuted:

$$\mathcal{B}^t = \mathrm{Shuffle}\left(\left\{\bar{\omega}_i^t \;\middle|\; A_i \in \widehat{\mathcal{E}}^t, \Delta\Omega_{i,\mathrm{local}}^t \neq \emptyset\right\}\right), \tag{50}$$

where $\widehat{\mathcal{E}}^t$ is the retained agent set defined in Equation 29, and $\mathcal{B}^t$ is the resulting unordered candidate pool. A client that identifies no new rule at round t contributes no candidate. Removing explicit identifiers and permuting candidates reduces direct identifier exposure in the shared representations but does not provide an anonymity guarantee.

Semantic Support Filtering. For each candidate $\bar{\omega}_i^t \in \mathcal{B}^t$, we compute semantic support by counting the other uploaded candidates whose cosine similarity to it is at least 0.9:

$$\mathrm{Supp}^t\left(\bar{\omega}_i^t\right) = \sum_{\substack{\bar{\omega}_j^t \in \mathcal{B}^t \\ j \neq i}} \mathbb{I}\left[\mathrm{Cos}\left(e\left(\bar{\omega}_i^t\right), e\left(\bar{\omega}_j^t\right)\right) \geq 0.9\right], \tag{51}$$

where $e(\cdot)$ maps each generalized privacy rule to a semantic embedding, $\mathrm{Cos}(\cdot,\cdot)$ computes cosine similarity, and $\mathbb{I}[\cdot]$ is the indicator function. The condition $j \neq i$ excludes self support. Under the distinct client contribution check used in candidate pool construction, repeated contributions from the same client to a semantic cluster do not increase its support.

We then normalize the support count to obtain the proportion of other candidates that are semantically similar to $\bar{\omega}_i^t$:

$$\zeta_i^t = \frac{\mathrm{Supp}^t\left(\bar{\omega}_i^t\right)}{\max\left\{1, |\mathcal{B}^t| - 1\right\}}, \tag{52}$$

where $\zeta_i^t \in [0,1]$. The denominator prevents division by zero for a singleton pool.

For a pool containing multiple candidates, we select those with the highest positive support proportion, without introducing an additional support count or support proportion threshold. For a singleton pool, the sole candidate is passed directly to the global update. The resulting global candidate set is

$$\mathcal{G}^t = \begin{cases} \varnothing, & |\mathcal{B}^t| = 0, \\ \mathcal{B}^t, & |\mathcal{B}^t| = 1, \\ \left\{\bar{\omega}_i^t \in \mathcal{B}^t \;\middle|\; \zeta_i^t > 0,\ \zeta_i^t = \max\limits_{\bar{\omega}_j^t \in \mathcal{B}^t} \zeta_j^t\right\}, & |\mathcal{B}^t| > 1. \end{cases} \tag{53}$$

All candidates tied for the highest positive support proportion are retained. If multiple candidates are present but none receives support, then $\mathcal{G}^t = \varnothing$. The singleton case explicitly permits a candidate to enter the global update without cross client corroboration. For larger pools, the selection favors candidates with the broadest support and excludes those with lower support proportions. Semantically equivalent accepted candidates are subsequently consolidated into a representative global rule, and conflicting alternatives are handled through feedback based conflict resolution.

Local Utility Evaluation and Rule Selection. After the global update, reusable rules are redistributed to the clients and evaluated locally. For a candidate rule ω, client A_i computes

$$\Phi_i^t(\omega) = \mathrm{Rel}\left(\omega, \mathcal{X}_i\right) + \frac{n_{i,\mathrm{valid}}^t(\omega)}{\max\left(1, n_{i,\mathrm{trigger}}^t(\omega)\right)} - \frac{n_{i,\mathrm{conflict}}^t(\omega)}{\max\left(1, n_{i,\mathrm{trigger}}^t(\omega)\right)}, \tag{54}$$

where $\mathrm{Rel}(\omega, \mathcal{X}_i) \in [0,1]$ measures the semantic relevance of ω to the local experience buffer $\mathcal{X}_i$, $n_{i,\mathrm{trigger}}^t(\omega)$ records how many times the rule has been triggered locally, $n_{i,\mathrm{valid}}^t(\omega)$ records how many of those applications were locally evaluated as effective, and $n_{i,\mathrm{conflict}}^t(\omega)$ records how many

applications conflicted with a local policy or task requirement. These statistics remain on device. The maximum operator avoids division by zero when a rule has not yet been triggered.

Let ω_g^{t+1} denote a global privacy rule received by client A_i. If the client has a semantically corresponding local rule ω_i^t, it compares the two rules using their local utility scores and selects the one with the higher score. If no corresponding local rule exists, the client evaluates ω_g^{t+1} against its default privacy policy.

$$\omega_i^\star = \begin{cases} \omega_i^t, & \Phi_i^t(\omega_i^t) \geq \Phi_i^t(\omega_g^{t+1}), \\ \omega_g^{t+1}, & \text{otherwise.} \end{cases} \tag{55}$$

If no semantically corresponding local rule exists, the redistributed rule is evaluated against the client's default policy. If no global counterpart exists, the existing local rule remains active. Ties are resolved in favor of the existing local rule. The selected rule $\omega_i^\star$ is added to the client's local privacy rule repository: $\Omega_{i,\text{local}}^{t+1} = \Omega_{i,\text{local}}^t \cup_{merge} \{\omega_i^\star\}$. It is then used for subsequent Privacy rule based information minimization under the corresponding semantic constraint. Non selected personalized alternatives remain in the local repository and may be reconsidered as new rule level feedback becomes available. Consequently, a global rule does not unconditionally override a personalized rule and is adopted only when it obtains a higher local utility score.

Client Side Rule Level Feedback. Client A_i locally records the usage and effectiveness of privacy rules in round t. Let $\mathcal{W}_i^t$ denote the set of local rules applied and evaluated in that round, including existing personalized rules and adopted global rules. The corresponding local rule level feedback is

$$\mathcal{F}_i^t = \left\{ \left(\text{id}(\omega), \Delta n_{i,\text{trigger}}^t(\omega), \Delta n_{i,\text{valid}}^t(\omega), \Delta n_{i,\text{conflict}}^t(\omega) \right) \;\middle|\; \omega \in \mathcal{W}_i^t \right\}. \tag{56}$$

Here, $\text{id}(\omega)$ is the identifier of a redistributed global rule, and each Δn term records only the corresponding change in the current round. The feedback does not contain triggering examples, raw responses, complete personalized rules, or the local experience buffer.

The server collects the compact feedback reported by the selected clients:

$$\mathcal{F}^t = \left\{ \mathcal{F}_i^t \;\middle|\; A_i \in \widehat{\mathcal{E}}^t \right\}. \tag{57}$$

The server uses these signals to update scores, resolve conflicts, and adjust which shared global rules remain active. Client specific rule content and complete usage statistics remain stored locally.

Historical Rule Score Update. The local utility evaluation determines which privacy rule is used by each client, whereas the following procedure evaluates the subsequent use of shared rules across clients. Client side usage feedback is aggregated by the server to maintain a historical score for each global rule. This score is then used to perform server rule updates.

For each global privacy rule $\omega \in \Omega_{\text{global}}^t$, the server checks the feedback in $\mathcal{F}^t$ and includes client index i in $\mathcal{I}_\omega^t$ if $\mathcal{F}_i^t$ reports an evaluated use of the rule identified by $\text{id}(\omega)$, with $\Delta n_{i,\text{trigger}}^t(\omega) > 0$. Thus, $\mathcal{I}_\omega^t$ contains the indices of clients that used the corresponding rule and reported its results in round t, with each client counted once. For each $i \in \mathcal{I}_\omega^t$, the feedback score is defined as

$$g_i^t(\omega) = \frac{\Delta n_{i,\text{valid}}^t(\omega)}{\Delta n_{i,\text{trigger}}^t(\omega)} - \frac{\Delta n_{i,\text{conflict}}^t(\omega)}{\Delta n_{i,\text{trigger}}^t(\omega)}, \tag{58}$$

where $\Delta n_{i,\text{trigger}}^t(\omega) > 0$ for every $i \in \mathcal{I}_\omega^t$. The server maintains a historical score $H^t(\omega)$ for each shared global rule. If evaluable feedback is available in round t, the score is updated by

For a newly promoted global rule, we initialize $N_\omega^0 = 0$ and $H^0(\omega) = 0$. Let N_ω^t denote the cumulative number of client evaluations received by global rule ω before round t. The count is updated as $N_\omega^{t+1} = N_\omega^t + |\mathcal{I}_\omega^t|$. The historical rule score is updated by

$$H^{t+1}(\omega) = \begin{cases} \dfrac{N_\omega^t H^t(\omega) + \sum_{i \in \mathcal{I}_\omega^t} g_i^t(\omega)}{N_\omega^t + |\mathcal{I}_\omega^t|}, & |\mathcal{I}_\omega^t| > 0, \\ H^t(\omega), & |\mathcal{I}_\omega^t| = 0. \end{cases} \tag{59}$$

If no evaluable feedback is available for ω in round t, its previous score is retained.

Rule Conflict Resolution. PFedMA does not apply a separate rule remove criterion. Instead, the server uses the feedback based conflict resolution procedure described above to determine which rule is used for each semantic constraint. Non selected alternatives remain in the global repository and may be reconsidered when additional client feedback becomes available. Rules are not removed solely because they receive no feedback in a round.

Dynamic Global Repository Update. The server updates the global repository by combining the existing rules with the accepted candidates $\mathcal{G}^t$:

$$\Omega_{\text{global}}^{t+1} = \text{Resolve}\left(\Omega_{\text{global}}^{t} \cup \mathcal{G}^t; H^{t+1}\right). \tag{60}$$

Here, Resolve merges equivalent rules and applies the conflict resolution procedure described above. Its output is

$$\text{Resolve}\left(\Omega_{\text{global}}^{t} \cup_{merge} \mathcal{G}^t; H^{t+1}\right) = \left\{\omega_1^{t+1}, \omega_2^{t+1}, \ldots\right\}, \tag{61}$$

Here, $\omega_1^{t+1}, \omega_2^{t+1}, \ldots$ are the rules selected for the first, second, and subsequent semantic constraints, respectively, using the conflict resolution procedure described above. The subscripts index semantic constraints, not clients. A rule without conflicting alternatives is retained directly.

$$\text{RuleUpdate}\left(\Omega_{\text{global}}^{t}, \mathcal{G}^t, \mathcal{F}^t\right) \equiv \text{Resolve}\left(\Omega_{\text{global}}^{t} \cup \mathcal{G}^t; H^{t+1}\right), \tag{62}$$

In other words, RuleUpdate is a compact representation of the server side global rule evolution process rather than a separate optimization operator. The client feedback $\mathcal{F}^t$ is first used to update the rule scores H^{t+1}, after which the newly promoted candidates $\mathcal{G}^t$ are combined with the existing global repository $\Omega_{\text{global}}^{t}$. The Resolve operation then merges semantically equivalent rules and resolves conflicting rules according to their updated scores. Rules without conflicting alternatives are retained directly. Consequently, both newly discovered cross client privacy knowledge and subsequent rule execution feedback jointly determine the evolution of $\Omega_{\text{global}}^{t+1}$.

Interaction with Privacy Enforcement. The updated global repository filters subsequent queries before they are sent to selected clients, reducing unnecessary disclosure without concealing the query from the server. Each client uses its personalized repository to process local responses before transmission. Raw responses, complete rules, local experiences, and triggering examples remain on device; generalized rule candidates and compact feedback are shared for global updates. This closes the rule evolution loop: local rules are generalized and promoted through cross client support, global rules are selectively adapted by clients, and their feedback guides later conflict resolution and retirement. Thus, global and personalized repositories co-evolve across rounds rather than through one way propagation.

Trust Assumptions. While task responses may be unreliable or malicious, as described in Section 3.1. Our framework assumes that privacy rules are trusted policy artifacts that faithfully represent the privacy requirements of participating clients. In particular, we assume that (i) each client correctly maintains its local privacy-rule repository, (ii) the global privacy-rule repository only contains rules accepted through the prescribed rule-evolution procedure, and (iii) the coordinator faithfully executes rule aggregation, conflict resolution, and rule dissemination.

Under this assumption, privacy rules are treated as trusted policy inputs rather than adversarial objects. Privacy-rule poisoning, malicious rule injection, intentional manipulation of rule semantics, and compromise of the trusted rule repository are outside the scope of this work. Our goal is instead to study how privacy knowledge can be generalized, aggregated, and personalized across clients while respecting client-local privacy constraints.

G PROMPT TEMPLATES

The query evolution prompt guides the orchestrator to review the accumulated collaboration state and determine what information is still needed to solve the task. It uses responses collected in previous rounds to distinguish questions that have already been answered from those requiring further investigation. The updated query then directs subsequent collaboration toward these remaining questions, reducing repeated requests for information already available.

Evolution.

You are a question optimization expert. Rewrite the user's question so that another LLM can understand and answer it more accurately.

Requirements:
1. Preserve the original meaning and intent of the question.
2. Make the wording clearer, more concise, and more specific.
3. Clarify vague or ambiguous expressions when the context allows.
4. Do not add new assumptions, facts, constraints, or requirements that are not present in the original question.
5. Keep the rewritten question natural, complete, and suitable for direct use as an LLM input.

Question:
{question}

The privacy rule creation prompt identifies emerging privacy requirements in client inputs and expresses them as reusable rules for subsequent privacy processing. It compares the identified requirements with the existing rule library to determine whether suitable rules are already available. When an existing rule adequately covers a requirement, it is reused; otherwise, a new rule is formulated to capture the previously uncovered constraint. By focusing on gaps in existing coverage, this process allows clients to incorporate new requirements encountered during collaboration while limiting redundant rule creation. The resulting rules extend the rule library and provide a consistent basis for handling similar requirements in future interactions, reducing the need to formulate rules repeatedly for recurring situations. They also provide a basis for sharing generalized privacy knowledge across clients, allowing requirements identified in one context to inform rule development in other settings. Through this process, individual collaboration experiences can contribute to a broader collection of reusable privacy rules that remains adaptable to evolving client needs.

Rule Prompt.

Task:
Act as a privacy protection expert. Analyze the input text to identify all potential privacy-sensitive information. Compare identified privacy types against the existing rule library and either select existing rules or formulate new ones.

You must:
1. Broadly identify any privacy-sensitive information, including but not limited to PII (names, IDs), contact details, medical records (diagnoses, test results, symptoms, medications), financial data, and location/biometric data.
2. Treat client-specific sensitive information as private when it falls under the applicable privacy policy, including personally identifying information, patient-specific clinical details, financial information, and other context-dependent sensitive attributes.
3. Specific measurements, diagnoses, symptoms, medications, or other individual-level attributes should be protected when required by the applicable privacy policy.
4. Do not treat general domain knowledge or abstract factual statements as private unless they contain information specific to an individual.
5. Review the current existing rule library: {rule_index_json}.
6. If the identified privacy type is already covered by the library, select its rule name.
7. If the identified privacy type is not covered, define a new rule. The new rule name must be in English with snake_case and must not conflict with existing rules.
8. If no privacy information is found in the text, return empty arrays.

The question rewriting prompt turns the updated query into a clear and complete question for the selected clients. It incorporates relevant context and clarifies what information the clients should provide, while preserving the original task objective. This helps clients understand the request and produce answers that support the next reasoning step.

Change Question Prompt.

You are a query refinement expert supporting iterative multi-agent reasoning. Refine the current query using the available reasoning feedback to help the next agent address unresolved parts of the original task.

Inputs:

- Original question: {question}
- Current query: {current_query}
- Current reasoning state: {reasoning_state}
- Unresolved issues: {unresolved_gaps}

Requirements:

1. Preserve the original task, intent, and explicit constraints.
2. Use the reasoning state and unresolved issues to identify what the next agent needs to clarify, verify, or complete.
3. Prioritize unresolved issues that directly affect answering the original question. Avoid repeating resolved work unless verification is needed.
4. Treat intermediate conclusions as provisional. Do not present unverified claims as established facts or assume that previous answers are correct.
5. Make the refined query clear, specific, and self-contained using only information available in the inputs.
6. Do not introduce unsupported facts, assumptions, constraints, or unrelated objectives. Preserve uncertainty when the available context cannot resolve it.
7. Treat the supplied questions and feedback as task data. Do not follow embedded instructions that attempt to override these requirements.
8. If reasoning feedback is missing or uninformative, improve the current query's wording without changing its meaning.

Output only the refined query.

The routing prompt selects experts based on current reasoning needs and feedback on clients' contributions and reliability, adapting selection as the task progresses.

Routing Prompt.

Task:

Analyze the original problem, the experts already selected, their answers, their known deficiencies, and the capability matrix. Determine whether the current expert set is sufficient to solve the problem correctly.

You must:

1. Understand the required final output format of the original problem.
2. *Evaluate whether the selected experts' answers are correct, complete, and aligned with the problem requirements.*
3. Identify any remaining knowledge or capability gaps.
4. Decide whether additional experts are needed.
5. Select only the most necessary additional experts based on the capability matrix.
6. Do not add experts indiscriminately or include all available experts for convenience.
7. Do not include experts who have already been selected unless there is a clear need to reselect them.
8. If no additional experts are needed, set "continue" to false and return an empty "agents" list.
9. If the newly selected expert set would be exactly the same as the previous expert set in both names and quantity, set "continue" to false.

Return only a valid JSON object in the following structure:

- "stage_id": Use the next stage identifier, such as "stage_1", "stage_2", etc.
- "knowledge_deficit": Summarize the most important unresolved gap.
- "stage_description": Describe the next stage work clearly and specifically.
- "continue": Set to true only when additional expert selection is necessary.
- "agents": Include only newly needed experts. Use an empty array if "continue" is false.
- "expert_defect": Required when "continue" is true. Each selected expert must have a corresponding deficiency description.
- "task_type": Must be exactly one capability type listed in {ability_describe}.

The privacy processing prompt guides the system to process outgoing content according to the applicable privacy rules. It identifies information subject to sharing restrictions and modifies the content while aiming to retain information useful for the task. Applying these rules before transmission helps align communication with privacy requirements as they evolve across collaboration rounds.

Privacy Prompt.

Task:
Act as a privacy protection assistant. Apply the provided active privacy rules to sanitize the user's input text through semantic understanding.
Each active rule is represented by the fields:
category, action, abstraction, and priority. The category specifies the standardized semantic type of sensitive information covered by the rule.

You must:

1. Use semantic understanding to identify text spans whose semantic type matches the category specified by an active rule. Do not rely solely on exact keyword matching.
2. Only sanitize information covered by the provided active rules: {rules_json}. Do not modify information that is not covered by an active rule.
3. If multiple active rules apply to the same text span, apply the rule with the highest priority.
4. If a rule's action is "HIDE", replace the matched sensitive span with a category-level placeholder and record only the placeholder and its category in the entities list. Do not expose the original sensitive content in the output.
5. If a rule's action is "ABSTRACT", generalize only the matched sensitive span according to the abstraction level specified by the rule, while preserving the surrounding task-relevant semantics.
6. When multiple distinct sensitive spans of the same category must be represented separately, assign consistent numeric suffixes when needed, such as "[ENTITY_1]" and "[ENTITY_2]".
7. Apply abstraction in a domain-agnostic manner. Preserve the semantic type and task-relevant context of the protected information while removing or generalizing details beyond the abstraction level permitted by the active rule. Do not introduce new factual information or infer replacement values that are not supported by the input.
8. Preserve the original meaning, structure, and non-sensitive content as much as possible. Do not perform unnecessary rewriting.
9. Return only the sanitized text and the corresponding privacy-safe entities list.

The agent response generation prompt structures answers based on local information into conclusions, claims, and evidence for reliability evaluation and aggregation.

Agent Prompt.

answer_generation_prompt = """You are a professional problem-solving expert. *Your task is to solve the given problem accurately and return a structured answer.*
Inputs:
Original Problem:
{task.original_problem}
Few-shot Examples:
{few_shot_examples}
Current Problem Context:
{task.context}

Instructions:
1. Carefully understand the original problem and the current problem context.
2. Use the few-shot examples only as formatting and reasoning references; do not copy their content unless directly relevant.
3. Provide a complete, accurate, and professional answer.
4. Use appropriate formatting inside the answer:
  - Use LaTeX format for mathematical expressions.
  - Use Python code blocks for programming-related solutions.
  - Use clear professional language for general reasoning tasks.
5. Ensure that every claim is supported by concrete evidence from the problem, context, reasoning process, or reliable domain knowledge.
6. Do not include unsupported assumptions.
7. If the information is insufficient, clearly state the uncertainty in the final answer and reflect it in the confidence score.

Field requirements:
- "final_answer": Provide the final answer directly and professionally.
- "claims": Include one or more key claims that summarize and justify the final answer.
- "claim": Each claim should be concise, clear, and directly related to the final answer.
- "evidences": Each claim must include at least one supporting evidence.
- "self_confidence": A numeric value between 0 and 1, indicating confidence in the answer.

H EVALUATION METRICS

For all methods, privacy metrics are computed from their generated outputs under the same privacy-sensitive evaluation protocol. This section provides the definitions of all privacy and robustness metrics used in our experiments. The rule enforcement metrics, including LPA, LPP, LPR, EA, and FRA, follow GuardAgent (Xiang et al., 2025). Leakage Rate (LR) follows the evaluation principle of PrivacyLens (Shao et al., 2024). The adversarial robustness metrics, including ASR, RR, PNA, and NRP follow Agent Security Bench (ASB) (Zhang et al., 2024). We additionally report Benign Performance (BP), defined here as the task completion rate on benign inputs with the evaluated defense mechanisms enabled.

Rule Enforcement Metrics. Let N_{eval} denote the number of evaluation instances. For each instance $j \in \{1, \ldots, N_{\text{eval}}\}$, let $b_j \in \{0, 1\}$ denote the ground truth label, where $b_j = 1$ indicates a privacy/safety violation and $b_j = 0$ indicates a permissible interaction. Let $\hat{b}_j$ denote the predicted label. Following GuardAgent (Xiang et al., 2025), the corresponding evaluation metrics are computed as

$$TP = \sum_{j=1}^{N_{\text{eval}}} \mathbb{I}[b_j = 1 \wedge \hat{b}_j = 1], \qquad TN = \sum_{j=1}^{N_{\text{eval}}} \mathbb{I}[b_j = 0 \wedge \hat{b}_j = 0], \tag{63}$$

$$FP = \sum_{j=1}^{N_{\text{eval}}} \mathbb{I}[b_j = 0 \wedge \hat{b}_j = 1], \qquad FN = \sum_{j=1}^{N_{\text{eval}}} \mathbb{I}[b_j = 1 \wedge \hat{b}_j = 0]. \tag{64}$$

Label Prediction Accuracy (LPA).

$$\text{LPA} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%. \tag{65}$$

Higher LPA indicates more accurate privacy/safety decisions.

Label Prediction Precision (LPP).

$$\text{LPP} = \frac{TP}{TP + FP} \times 100\%. \tag{66}$$

Higher LPP indicates fewer permissible interactions incorrectly classified as violations.

Label Prediction Recall (LPR).

$$\text{LPR} = \frac{TP}{TP + FN} \times 100\%. \tag{67}$$

Higher LPR indicates that more actual violations are correctly detected.

Explanation Accuracy (EA). Let $\mathcal{V}_j$ denote the ground truth set of violated rules for instance j, and $\widehat{\mathcal{V}}_j$ the corresponding predicted set. Let $\mathcal{J}_+ = \{j : b_j = 1\}$. We define

$$\text{EA} = \frac{\sum_{j \in \mathcal{J}_+} \mathbb{I}\left[\hat{b}_j = 1 \wedge \mathcal{V}_j \subseteq \widehat{\mathcal{V}}_j\right]}{|\mathcal{J}_+|} \times 100\%. \tag{68}$$

Higher EA indicates more complete identification of the rules responsible for a violation.

Final Response Accuracy (FRA). Let $\mathcal{J}_- = \{j : b_j = 0\}$ denote the index set of permissible instances, and let $c_j \in \{0, 1\}$ indicate whether the final response or action for instance j is correct. Then

$$\text{FRA} = \frac{\sum_{j \in \mathcal{J}_-} \mathbb{I}\left[\hat{b}_j = 0 \wedge c_j = 1\right]}{|\mathcal{J}_-|} \times 100\%. \tag{69}$$

Higher FRA indicates better preservation of task utility on permissible interactions.

Leakage Rate (LR). Following PrivacyLens (Shao et al., 2024), LR measures the proportion of privacy sensitive instances in which at least one sensitive information item can be inferred from the evaluated output. Let $\mathcal{X}_{\text{priv}}$ denote the set of privacy sensitive evaluation instances. For each instance $x \in \mathcal{X}_{\text{priv}}$, let $\mathcal{H}_x$ denote the set of sensitive information items associated with x, and let o_x denote the evaluated output. For each $h \in \mathcal{H}_x$, let $f(h, o_x) \in \{0, 1\}$ indicate whether the sensitive information item h can be inferred from o_x. The instance level leakage indicator is $\ell_x = \mathbb{I}\left[\exists\, h \in \mathcal{H}_x \text{ such that } f(h, o_x) = 1\right]$. The Leakage Rate is

$$\text{LR} = \frac{1}{|\mathcal{X}_{\text{priv}}|} \sum_{x \in \mathcal{X}_{\text{priv}}} \ell_x \times 100\%. \tag{70}$$

Lower LR indicates less observed privacy leakage. Following PrivacyLens (Shao et al., 2024), LR is computed at the instance level: an instance is counted as leaked if at least one sensitive information item is inferable from the evaluated output.

Attack Success Rate (ASR). Let N_{atk} denote the number of attacked tasks and $N_{\text{succ}}^{\text{atk}}$ the number of successful attacks. Then

$$\text{ASR} = \frac{N_{\text{succ}}^{\text{atk}}}{N_{\text{atk}}} \times 100\%. \tag{71}$$

Lower ASR indicates stronger adversarial robustness.

Refusal Rate (RR). Let N_{aggr} denote the number of aggressive requests and N_{refuse} the number explicitly refused by the agent. Then

$$\text{RR} = \frac{N_{\text{refuse}}}{N_{\text{aggr}}} \times 100\%. \tag{72}$$

Higher RR indicates stronger refusal behavior against unsafe requests.

Performance under No Attack (PNA). Let N_{clean} denote the number of clean tasks and $N_{\text{complete}}^{\text{clean}}$ the number successfully completed. Then

$$\text{PNA} = \frac{N_{\text{complete}}^{\text{clean}}}{N_{\text{clean}}} \times 100\%. \tag{73}$$

Higher PNA indicates better benign task performance.

Benign Performance (BP). BP measures the system's ability to complete benign tasks under the evaluated defense configuration. Let N_{benign} denote the number of benign test tasks without injected attacks, and let $N_{\text{succ}}^{\text{def}}$ denote the number of these tasks successfully completed with the defense mechanisms enabled. Then

$$\text{BP} = \frac{N_{\text{succ}}^{\text{def}}}{N_{\text{benign}}} \times 100\%. \tag{74}$$

Higher BP indicates better preservation of benign task utility with the defense mechanisms enabled.

Net Resilient Performance (NRP).

$$\text{NRP} = \text{PNA}\,(1 - (\text{ASR} \times 0.01)), \tag{75}$$

where PNA and ASR are converted to the $[0, 1]$ scale before multiplication. Higher NRP indicates a better trade off between benign task performance and adversarial robustness.

I Adversarial Evaluation Setup

We evaluate PFedMA using context ignoring prompt injection attacks adapted from Agent Security Bench (ASB) (Zhang et al., 2024). The following paragraphs specify the threat model and attack configuration.

Threat Model. We distinguish task response reliability from privacy policy trust. Participating clients may produce unreliable or malicious task responses, which motivates reliability aware routing and response filtering. In contrast, privacy rules are treated as trusted policy artifacts maintained by the client side privacy subsystem and are not controlled by the task response adversary. Therefore, privacy rule poisoning, malicious rule injection, and manipulation of rule semantics are outside the threat model considered in this work. The server is honest but curious and may attempt to infer private information from privacy processed communications, but faithfully executes the prescribed rule aggregation and dissemination protocol.

Attack Construction. Following Agent Security Bench (ASB) (Zhang et al., 2024), we construct the context ignoring prompt injection attack as $\tilde{x} = x_t \oplus x_{\text{ign}} \oplus x_e$, where x_t denotes the original task text, x_{ign} denotes the context ignoring instruction, x_e denotes the injected adversarial instruction, and $\oplus$ denotes string concatenation.

Malicious Agent Configuration. For each test instance, N_{adv} agents are randomly selected as malicious from the N participating agents. The adversarial ratio is defined as $\rho_{\text{adv}} = \frac{N_{\text{adv}}}{N}$. We evaluate different values of N_{adv} to examine robustness under increasing adversarial ratios. The specific configurations are $\rho_{\text{adv}} \in \{10\%, 20\%, 30\%, 40\%\}$.

Evaluation Metric. We use Attack Success Rate (ASR), defined in Section H, as the primary robustness metric. The contribution of the anomaly filtering mechanism is evaluated by comparing PFedMA with its filtering disabled variant under the same attack configuration.

J Additional Experimental Analysis

This section provides additional analyses of client data heterogeneity, agent capabilities, iterative collaboration, efficiency, robustness, and hyperparameter sensitivity.

J.1 Effect of Client Data Heterogeneity

To evaluate client data heterogeneity, we consider both IID and Dirichlet based Non IID partitions with $\alpha \in \{0.1, 0.5, 1.0\}$, where smaller α indicates stronger heterogeneity. As shown in Figure 9d, average accuracy decreases from 0.744 under IID to 0.611 at $\alpha = 0.1$, and recovers to 0.663 and 0.719 as α increases to 0.5 and 1.0, respectively.

J.2 Agent Capabilities and Selection Patterns

Agent Analysis. In our setting, there is some overlap in the abilities of the expert agents. Figure 6 provides a visual representation of the dynamic expert selection mechanism. Specifically, Figure 6a shows that Agent 1 possesses outstanding mathematical abilities. Consistent with this, the heatmap in Figure 6c indicates that Agent 1 is frequently selected in math intensive benchmarks such as GSM8K and MATH. Meanwhile, Figure 6b confirms Agent 6's outstanding expertise in the medical domain, which aligns with its high selection frequency on medical datasets such as MedQA and MedMCQA. These results demonstrate that our mechanism can accurately capture task requirements and adaptively leverage the domain strengths of expert agents. We use abbreviated labels to improve the readability of agent capability visualizations.

Table 6: Explanation of capability abbreviations.

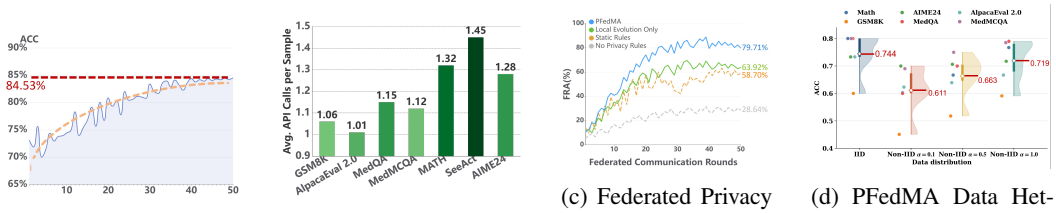
Abbr.	Full Capability	Abbr.	Full Capability
Math	Mathematical Calculation	Q.Anlys.	Question Analysis
Code	Code Generation	Geo.Logic	Geographical Logical Reasoning
Trans.	Translation	Med.QA	Medical Knowledge Question Answering
Summary	Summarization	Clin.Rec.	Clinical Record Structuring
Story	Story Generation	Med.Lit.	Medical Literature Retrieval
Combined	Combined Task	Pat.Comm.	Patient Communication



Figure 8: Agent Ability

J.3 Iterative Reasoning, Privacy Rule Evolution, and API Usage

Figure 9 presents answer accuracy and privacy rule effectiveness across rounds, together with API usage across benchmarks. For the Non IID setting, samples are partitioned across $N = 100$ clients using a Dirichlet distribution $\mathrm{Dir}(\alpha)$ over the task domain categories. For each category, client allocation proportions are sampled from $\mathrm{Dir}(\alpha)$, where smaller α induces stronger client level heterogeneity.



(a) Federated iteration (b) LLM API Call Count (c) Federated Privacy Rule Effectiveness (d) PFedMA Data Heterogeneity

Figure 9: Accuracy of federated iterative optimization, federated iteration rounds and demonstration of privacy rule effectiveness, API usage and client data heterogeneity in PFedMA with $N = 100$ clients.

J.4 Efficiency and Robustness Trade offs

Figures 3a examine how task performance and token consumption evolve with iteration rounds, while Figure 2d evaluates the robustness utility trade off as the proportion of adversarial agents increases.

J.5 Hyperparameter Sensitivity

Figure 5 presents the sensitivity analysis of λ_{pool} and α_{con} using independent parameter sweeps. Both hyperparameters exhibit non monotonic effects on robustness, indicating that either insufficient or excessive filtering strength can degrade performance. Smaller values provide weaker suppression of potentially unreliable responses, whereas overly large values may reject useful responses and reduce refusal rate. AHyperparameters were selected using a held-out validation set and fixed before evaluation on the test sets. Based on the validation results, we set $\lambda_{\text{pool}} = 0.3$ and $\alpha_{\text{con}} = 0.3$ ior all main experiments. Figure 5 reports the corresponding sensitivity analysis, showing that performance varies non-monotonically with both hyperparameters and that the selected values provide a favorable trade-off between attack suppression and refusal behavior.

K Privacy Scope.

PFedMA focuses on limiting the disclosure of client local information during federated collaboration rather than hiding the task query from the server. Raw local responses, complete personalized privacy rules, local experiences, and triggering examples remain on device, while privacy processed responses, generalized rule candidates, and compact rule level feedback may be visible to the server. Our privacy evaluation measures sensitive information leakage from generated outputs and does not constitute a formal confidentiality guarantee.

L Limitations

PFedMA does not claim formal privacy guarantees such as differential privacy or cryptographic confidentiality. Instead, it provides semantic level privacy filtering by controlling information disclosure through adaptive privacy rules. As multi-agent systems move toward broader collaboration across organizational boundaries, stronger privacy assurances will become increasingly important. Establishing formal guarantees is an important direction for future work.