SPEND WHERE IT MATTERS: FINE-GRAINED PRIVACY ALLOCATION FOR EDGE-CLOUD REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Privacy-sensitive multi-step reasoning must exploit the capabilities of cloud large language models (LLMs) without exposing private knowledge stored on trusted edge devices. Local-only execution protects privacy but limits reasoning quality, whereas existing edge–cloud approaches typically make coarse request-level decisions that cannot accommodate the heterogeneous privacy and quality requirements of individual reasoning steps. We formulate this problem as fine-grained privacy allocation over a reasoning graph and introduce *Operational Privacy Dependency* to characterize step-specific privacy–utility frontiers. We further derive a lower bound on the privacy cost required to achieve a target reasoning distortion and establish an adaptive end-to-end privacy guarantee. Building on this theory, we propose Swim, which constructs privacy-valid actions, analytically estimates their state-dependent quality–privacy operating points, and performs frontier-aware online allocation under a global privacy budget. Across six dataset–model settings, Swim retains 95.7% of Cloud-Only F1, improves F1 by 5.6% over the strongest non-Cloud-Only baseline, and reduces privacy-leakage AUC and MI by 18.3% and 39.1%, respectively. Code is available at `https://anonymous.4open.science/r/Swim-4EB3/`.

1 INTRODUCTION

Knowledge-intensive reasoning often requires large language models (LLMs) to combine parametric reasoning capabilities with external evidence (Lewis et al., 2020; Gao et al., 2023). Such reasoning is increasingly deployed in privacy-sensitive domains such as healthcare, finance, and enterprise question answering (Xiong et al., 2024; Choi et al., 2025; Packowski et al., 2024). In these applications, however, the evidence required for reasoning may include personal records, enterprise documents, or proprietary knowledge that cannot be directly exposed to an untrusted cloud. Existing privacy-preserving reasoning systems therefore seek to protect user queries, external evidence, or intermediate representations during cloud execution (Cheng et al., 2025; Pang et al., 2024).

A conservative solution is to keep private knowledge and model execution entirely on the trusted edge. Recent advances in mobile edge intelligence and sparse-model deployment have improved the feasibility of running language models on resource-constrained devices (Qu et al., 2025; Yi et al., 2025). Nevertheless, edge SLMs remain limited by computation, and memory on complex reasoning tasks. Fully local execution therefore protects privacy but sacrifices cloud-level reasoning quality.

Edge–cloud collaboration offers an alternative by coordinating local SLMs and cloud LLMs rather than reproducing the full cloud model at the edge. Existing studies employ mechanisms such as secure cloud guidance, privacy-preserving retrieval, cost-aware task decomposition, and sensitivity-aware routing (Zhang et al., 2024a; Li et al., 2025; Zhang et al., 2025; Narayan et al., 2025; Zhan et al., 2026) (see Appendix A for a broader discussion). These approaches demonstrate the potential of edge–cloud collaboration, but they are not designed to allocate a shared privacy budget across the interdependent steps of a reasoning process. Meanwhile, complex reasoning is increasingly organized through decomposed subtasks, interleaved retrieval and reasoning, or graph-structured intermediate states (Khot et al., 2022; Trivedi et al., 2023; Besta et al., 2024). A single collaboration decision for the entire request can therefore overlook substantial differences among reasoning steps.

Figure 1 reveals two coupled forms of heterogeneity in multi-step edge–cloud reasoning. First, reasoning steps within the same request can differ substantially in both their dependence on private

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

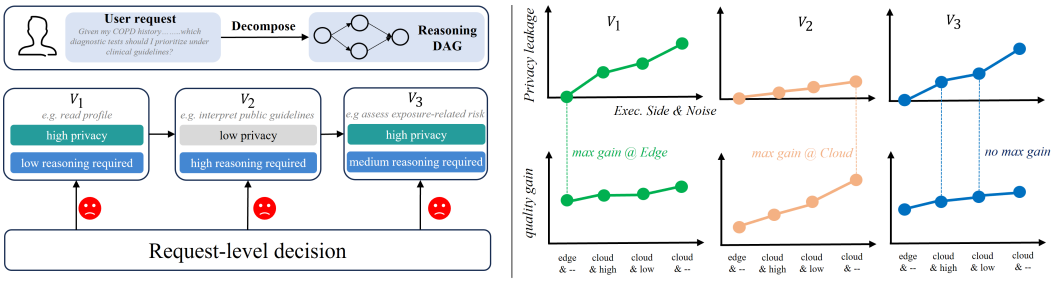

Figure 1: **Motivation for fine-grained privacy allocation.** *Left*: Request-level collaboration cannot capture the heterogeneous privacy needs of interdependent reasoning steps. *Right*: Preliminary analysis shows that the same privacy expenditure can yield markedly different reasoning gains across steps, motivating budgeted step-level allocation.

information and their demand for cloud reasoning: some steps are privacy-sensitive but require limited cloud assistance, whereas others can benefit greatly from cloud reasoning while relying primarily on public information. Second, the same level of privacy expenditure can yield markedly different quality gains across steps. Consequently, a single request-level collaboration decision must treat heterogeneous steps uniformly, which can either expose private information unnecessarily or fail to provide cloud assistance where it is most valuable.

To address these, we propose **Swim** (**S**pend **w**here **i**t **m**atters), a fine-grained privacy allocation framework for edge–cloud reasoning. Swim formulates privacy-constrained collaboration as an operating-point allocation problem over a reasoning graph, where decisions are coupled through both the evolving cloud-visible context and a shared global privacy budget. We introduce *Operational Privacy Dependency* to formalize the privacy–utility frontier of each reasoning step, derive a lower bound on the privacy required to attain a target reasoning distortion, and establish an adaptive end-to-end composition guarantee. Building on this foundation, Swim estimates state-dependent privacy–utility operating points and performs frontier-aware online allocation, balancing the expected quality return of a current action against its incremental privacy cost and the value of unresolved downstream reasoning. In this way, Swim directs privacy expenditure toward the steps where protected cloud assistance yields the greatest value while enforcing a global privacy constraint.

Our main contributions are summarized as follows:

- We formulate privacy-constrained edge–cloud reasoning as operating-point allocation over a graph, and consequently introduce *Operational Privacy Dependency* to characterize step-level privacy–utility frontiers and establish corresponding lower bounds and end-to-end privacy guarantees.
- We develop a theory-guided framework that constructs privacy-valid actions, estimates state-dependent operating points, and allocates a global privacy budget online.
- Across six dataset–model settings, Swim achieves 95.7% of Cloud-Only F1, exceeds the strongest non-Cloud-Only baseline by 5.6%, and reduces privacy-leakage AUC and MI by 18.3% and 39.1%, respectively. Ablation results further confirm the importance of analytic node-state estimation and future-aware privacy reservation.

2 Problem Formulation

We consider privacy-constrained knowledge-augmented reasoning in a fixed edge–cloud deployment. The trusted edge has exclusive access to private knowledge K_{pri} and hosts a resource-constrained small language model (SLM), while the honest-but-curious cloud hosts a more capable large language model (LLM). Public knowledge K_{pub} is externally accessible and may be used on either side.

Reasoning graph. Given a complex query Q, the trusted edge first produces a sanitized query $\widetilde{Q}$ that preserves the task intent while removing directly identifiable private content. A cloud-side planner

then constructs a reasoning DAG $G = (\mathcal{V}, \mathcal{E})$ from $\widetilde{Q}$ and public information. Once constructed, G is treated as fixed during node-level action allocation: each node $v_i \in \mathcal{V}$ represents a reasoning step, and $(v_i, v_j) \in \mathcal{E}$ indicates that v_j depends on the output of v_i. Swim focuses on selecting an execution and release action for each ready node.

Cloud-visible transcript and privacy. The cloud performs the requested computation faithfully but may attempt to infer private information from its observations. Let

$$\mathcal{T}_c = (M_t^{e\to c})_{t=1}^{T} \tag{1}$$

denote the cloud-visible transcript, where each $M_t^{e\to c}$ contains an edge-to-cloud release and any associated observable metadata that may depend on K_{pri}.

Two private knowledge bases $K_{\mathrm{pri}} \simeq K'_{\mathrm{pri}}$ are adjacent if they differ in one protected knowledge unit. We require the transcript distributions induced by adjacent knowledge bases to satisfy differential privacy; namely, for every measurable event $\mathcal{S}$,

$$\Pr[\mathcal{T}_c(K_{\mathrm{pri}}) \in \mathcal{S}] \le e^{\epsilon} \Pr\big[\mathcal{T}_c(K'_{\mathrm{pri}}) \in \mathcal{S}\big] + \delta.$$

A release that depends on previously unreleased private knowledge incurs a prescribed privacy cost, whereas computation based only on public or already released information incurs no additional privacy loss by post-processing. The privacy accountant composes the costs of all protected releases into an end-to-end privacy cost for the complete transcript.

Privacy-constrained reasoning allocation. Let $\pi \in \Pi(G)$ denote a dependency-respecting allocation policy that may adapt to previously observed reasoning outcomes. For each ready node v_i, the policy selects an action a_i specifying edge or cloud execution and, when necessary, a protected release configuration. Let $U(\pi)$ denote the utility of the final answer and let $P(\pi)$ denote the cumulative privacy cost reported by the accountant under a prescribed global δ. Given a global privacy budget B_{pri}, the allocation problem is

$$\begin{aligned} \max_{\pi\in\Pi(G)} \quad & \mathbb{E}[U(\pi)] \\ \text{s.t.} \quad & P(\pi) \le B_{\mathrm{pri}}. \end{aligned} \tag{2}$$

The decisions are coupled because upstream actions affect the contexts of downstream nodes, while all protected releases consume the same global privacy budget. The objective is therefore to allocate privacy across reasoning steps according to their state-dependent quality benefits and privacy costs.

3 Theoretical Foundations

We characterize the privacy-utility trade-off in edge-cloud reasoning and its implications for allocation. Conditioning each node on the preceding cloud-visible context isolates its incremental privacy cost, while prior disclosures are captured by end-to-end accounting.

3.1 Operational Privacy Dependency

For node v_i, condition on a realized cloud-visible context $X_i = x_i$. Let $Y_i(K_{\mathrm{pri}})$ denote its target answer under private knowledge K_{pri}, $d_i(\cdot,\cdot)$ a bounded distortion measure satisfying the triangle inequality, and $\mathcal{K}_i(x_i)$ the admissible private knowledge bases under x_i. We suppress the dependence on x_i when clear. Consider a cloud-assisted mechanism Γ_i producing $\widehat{Y}_i^c$. For fixed δ_i, let $\epsilon_i(\Gamma_i;\delta_i)$ be the smallest ϵ such that Γ_i satisfies (ϵ,δ_i)-DP over adjacent knowledge bases in $\mathcal{K}_i(x_i)$, conditioned on $X_i = x_i$.

Definition 1 (Operational Privacy Dependency)**.** For target distortion D, the operational privacy dependency of node v_i is

$$\rho_i(D;\delta_i) = \inf_{\Gamma_i:\, \sup_{K_{\mathrm{pri}}\in\mathcal{K}_i(x_i)} \mathbb{E}\left[d_i\left(Y_i(K_{\mathrm{pri}}),\widehat{Y}_i^c\right)\right] \le D} \epsilon_i(\Gamma_i;\delta_i).$$

Thus, $\rho_i(D;\delta_i)$ gives the minimum privacy budget needed to achieve distortion at most D under the fixed context. It is nonincreasing in D, and therefore characterizes a privacy–utility frontier. Further details are given in Appendix B.1.

3.2 Privacy-Utility Lower Bound

Let Δ_i be the maximum answer distortion between adjacent admissible private knowledge bases.

Theorem 2 (Privacy-Utility Lower Bound). *Conditioned on $X_i = x_i$, if a cloud-assisted mechanism for v_i satisfies (ϵ_i, δ_i)-DP and achieves expected distortion at most D_i uniformly over $\mathcal{K}_i(x_i)$, then*

$$D_i \geq \frac{(1-\delta_i)\Delta_i}{1+e^{\epsilon_i}}.$$

The theorem establishes an unavoidable privacy–utility trade-off: smaller target distortion requires a larger privacy budget. Appendix B.2 gives the proof and the corresponding lower bound on $\rho_i(D_i; \delta_i)$.

3.3 End-to-End Privacy Guarantee

A reasoning process may invoke the cloud repeatedly and adapt to earlier observations. Standard adaptive composition gives the following transcript-level guarantee.

Proposition 3 (Adaptive End-to-End Privacy). *If, conditioned on every preceding cloud-visible history, release $M_t^{e\to c}$ satisfies (ϵ_t, δ_t)-DP, then the complete cloud-visible transcript satisfies*

$$\left(\sum_t \epsilon_t, \sum_t \delta_t\right)\text{-DP}.$$

More generally, the accountant reports $P(\pi) = \epsilon_{\mathrm{acct}}(\pi; \delta)$, the end-to-end quantity constrained in Eq. (2). Appendix B.3 provides details.

3.4 Privacy-Constrained Reasoning Allocation

For action a_i, let $D_i(a_i)$ and $\epsilon_i(a_i)$ denote distortion and privacy cost. Definition 1 requires $\epsilon_i(a_i) \geq \rho_i(D_i(a_i); \delta_i)$; actions independent of K_{pri} have zero cost, while the accountant determines their contribution to $P(\pi)$.

The constrained problem in Eq. (2) admits the Lagrangian:

$$\mathcal{L}(\pi, \lambda) = \mathbb{E}[U(\pi)] - \lambda\left(P(\pi) - B_{\mathrm{pri}}\right), \qquad \lambda \geq 0. \tag{3}$$

Here λ is the marginal privacy price. Appendix B.3 derives the node-separable allocation condition, while Section 4 develops a receding-horizon allocation rule that accounts for shared-budget coupling and downstream opportunities.

4 Method

Swim performs fine-grained privacy allocation over the reasoning DAG by estimating privacy–quality operating points for each ready node and selecting actions under the evolving global privacy budget. The overall framework is illustrated in Figure 2. All allocation quantities use only local, public, or already accounted information, so action selection incurs no additional privacy cost.

4.1 Privacy-Valid Action Space

Let $\mathcal{R} = \{r_0, r_1, \ldots, r_L\}$ denote the available cloud release configurations, where r_0 transmits no private-information-dependent signal. The action space is $\mathcal{A} = \{a^e\} \cup \{a^c(r) : r \in \mathcal{R}\}$, where a^e denotes edge execution, $a^c(r)$ denotes cloud execution under configuration r, and $\mathcal{A}_i \subseteq \mathcal{A}$ contains the actions applicable to ready node v_i.

For $\ell \geq 1$, r_ℓ applies calibrated Gaussian noise with variance σ_ℓ^2 and satisfies $(\epsilon_\ell, \delta_\ell)$-DP, where $\epsilon_1 < \cdots < \epsilon_L$. Other DP mechanisms are supported by replacing the recovery factor in Eq. (6).

Thus, $\epsilon_i(a^e) = \epsilon_i(a^c(r_0)) = 0$ and $\epsilon_i(a^c(r_\ell)) = \epsilon_\ell$ for $\ell \geq 1$; the privacy accountant computes each action's incremental contribution online.

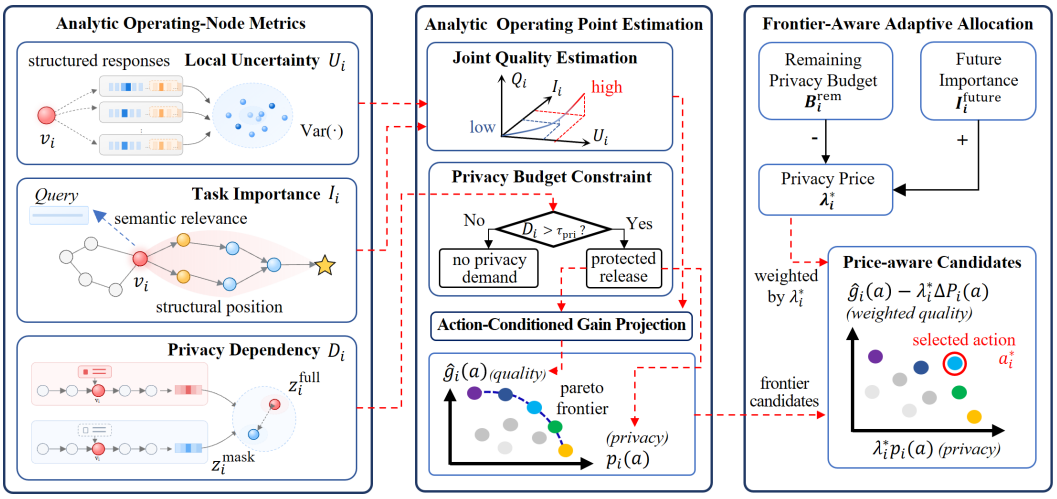

Figure 2: **Method overview.** For each ready node, we score privacy-valid actions locally and allocate the global privacy budget while reserving capacity for unresolved downstream nodes.

4.2 Analytic Operating-Node Metrics

Each configuration fixes its node-level DP guarantee, whereas its incremental global privacy cost depends on the current accountant state, as defined in Eq. (8). Trial cloud executions would themselves spend privacy, so a training-free model estimates every unexecuted candidate locally from three normalized statistics: local uncertainty U_i, task importance I_i, and privacy dependency D_i. The predicted gain $\widehat{g}_i(a)$ serves as a normalized local surrogate for the distortion reduction $D_i(a_i^e) - D_i(a)$ in Section 3.4.

Local uncertainty. For each ready node, the local SLM performs repeated stochastic executions under the same context. We measure dispersion among their local representations, weight output components by their use in downstream nodes, and normalize the result over the task DAG. The resulting $U_i \in [0, 1]$ is large when task-relevant local states disagree.

Task importance. Task importance combines DAG position with semantic relevance to the final answer. A frozen local encoder compares each node's public description with the original query, producing a semantic score $S_i \in [0, 1]$. Let $\mathcal{D}_{i \to t}$ contain the descendants of v_i that can reach answer node v_t. We discount their relevance by shortest-path distance $d(i, j)$ and normalize over the task DAG:

$$\widetilde{I}_i = \sum_{v_j \in \mathcal{D}_{i \to t} \cup \{v_i\}} \frac{S_j}{1 + d(i, j)}, \qquad I_i = \frac{\widetilde{I}_i}{\max\left\{\eta_I, \max_{v_k \in \mathcal{V}} \widetilde{I}_k\right\}}, \tag{4}$$

where the small constant $\eta_I > 0$ only prevents a zero denominator. Thus $I_i \in [0, 1]$.

Privacy dependency. We estimate private-information dependency with isomorphic twin DAGs: G^{full} uses complete local evidence, whereas G_i^{mask} replaces the private evidence of v_i with type-preserving placeholders. Let $\bar{z}_i^{\text{full}}$ and z_i^{mask} denote their corresponding local node representations. Because the twins otherwise share the same structure and prompts, their displacement isolates the effect of private evidence. We map it to $[0, 1]$ with a Gaussian kernel:

$$D_i^{\text{sem}} = 1 - \exp\left(-\frac{\left\|\bar{z}_i^{\text{full}} - z_i^{\text{mask}}\right\|_2^2}{2h^2}\right), \tag{5}$$

where $h > 0$ is fixed using public calibration embeddings. The score approaches zero when masking causes little change and one when it substantially changes the representation; we set $D_i^{\text{sem}} = 0$ if v_i receives no locally identified private evidence. The resulting analytic state is $Z_i^{\text{ana}} = (U_i, I_i, D_i^{\text{sem}})$.

4.3 Analytic Operating-Point Estimation

We define quality demand as the harmonic mean $Q_i = 2U_iI_i/(U_i + I_i)$, with $Q_i = 0$ when $U_i + I_i = 0$. Thus, $Q_i \in [0, 1]$ is high only when the node is both uncertain and important.

We treat dependencies at or below τ_{pri} as negligible, so the privacy-free cloud configuration r_0 can realize the full predicted gain. Otherwise, Gaussian variance σ_ℓ^2 determines the unrecovered private component under configuration r_ℓ. Combining this gate with private uncertainty $U_iD_i^{\text{sem}}$ gives

$$\widehat{g}_i(r_\ell) = Q_i \left[1 - \mathbb{I}\left(D_i^{\text{sem}} > \tau_{\text{pri}}\right) \frac{\sigma_\ell^2}{U_iD_i^{\text{sem}} + \sigma_\ell^2}\right]. \tag{6}$$

where $\mathbb{I}(\cdot)$ is the indicator function. Larger noise leaves more of the private component unrecovered; as σ_ℓ^2 decreases, the gain approaches Q_i. For r_0, setting $\sigma_0^2 \to \infty$ recovers no private-dependent component. Equation (6) uses only local statistics and the predefined mechanism variance.

We first apply the private-dependency gate. If $D_i^{\text{sem}} \le \tau_{\text{pri}}$, the privacy-free configuration r_0 provides the same predicted gain as all positive-privacy releases, which are therefore removed. Otherwise, Eq. (6) implements the *Action-Conditioned Gain Projection*, mapping each cloud action $a_i^c(r_\ell)$ to its predicted normalized distortion reduction $\widehat{g}_i(r_\ell)$. Together with $\widehat{g}_i(a_i^e) = 0$, each privacy-valid action $a \in \mathcal{A}_i$ is assigned an operating point $\mathcal{O}_i(a) = (\widehat{g}_i(a), p_i(a))$, where $p_i(a)$ is the accountant-derived incremental privacy cost in Eq. (8). We then remove Pareto-dominated actions and pass the remaining set $\mathcal{A}_i^\star$ to the allocator.

4.4 Frontier-Aware Adaptive Allocation

The analytic model supplies per-action gains; cumulative privacy cost additionally depends on the release history. We therefore couple it with the online accountant and reserve budget for unresolved descendants. To avoid counting a descendant repeatedly through several ancestors' aggregate importance scores, we sum each unresolved node's local semantic score once, discounted by its distance from v_i:

$$I_i^{\text{future}} = \sum_{v_j \in \mathcal{V}_i^{\text{unresolved}}} \frac{S_j}{1 + d(i, j)}, \tag{7}$$

where $\mathcal{V}_i^{\text{unresolved}}$ contains each unresolved descendant of v_i exactly once.

Let H_i be the current accountant state. The incremental cost of action a is

$$p_i(a) = \epsilon_{\text{acct}}\left(H_i \oplus a; \delta\right) - \epsilon_{\text{acct}}\left(H_i; \delta\right), \tag{8}$$

where $H_i \oplus a$ is the hypothetical state after adding all atomic releases associated with a.

We instantiate the marginal privacy price λ in Eq. (3) using a state-dependent receding-horizon surrogate. The remaining margin is $B_i^{\text{rem}} = B_{\text{pri}} - \epsilon_{\text{acct}}(H_i; \delta)$. With a small public stabilizer $\eta_B > 0$, the dynamic privacy price is

$$\lambda_i^\star = \frac{I_i^{\text{future}}}{\max\{B_i^{\text{rem}}, \eta_B\}}. \tag{9}$$

The price increases when unresolved importance is large relative to the remaining margin, raising the threshold for a current private release. The hard constraint below permits only zero-cost actions when $B_i^{\text{rem}} = 0$.

For the current ready node, the final action is selected as

$$a_i^\star = \underset{\substack{a \in \mathcal{A}_i^\star: \\ \epsilon_{\text{acct}}(H_i \oplus a; \delta) \le B_{\text{pri}}}}{\arg\max} \left[\widehat{g}_i(a) - \lambda_i^\star p_i(a)\right]. \tag{10}$$

Only $a_i^\star$ is executed, using configuration $r_i^\star = r(a_i^\star)$ and privacy expenditure $b_i^\star = p_i(a_i^\star)$. We then update the accountant and context, recompute the state of newly ready nodes, and update the future reserve and price. This receding-horizon rule accounts for downstream opportunities, while the hard accountant constraint prevents quality-model error from violating the global privacy budget.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETTINGS

Datasets. We evaluate Swim on three privacy-augmented QA datasets: *P-PubMedQA*, *P-FinQA* and *P-MultiHopRAG*, built by augmenting PubMedQA (Jin et al., 2019), FinQA (Chen et al., 2021), and MultiHopRAG (Tang & Yang, 2024) with private attributes, such as patient demographics and clinical history in P-PubMedQA. We jointly rewrite each question and reference answer to incorporate private evidence as a necessary reasoning step. An instance is retained only if the combined public and private evidence supports the answer, while removing the private evidence makes it incomplete or indeterminate. Appendix C.1 details the construction procedure and human validation.

Baselines. We compare Swim against seven representative baselines, including *Cloud-Only*, *RemoteRAG* (Cheng et al., 2025), *MINIONS* (Narayan et al., 2025), *CoGenesis* (Zhang et al., 2024a), *LSRP* (Zhang et al., 2025), *PRISM* (Zhan et al., 2026) and *Edge-Only*. More details about these baselines could be found in Appendix C.2.

Metrics and models. We report answer quality (F1 (Zhang et al., 2020), SM (Xu et al., 2025), and LaJ (Zheng et al., 2023)), empirical leakage (attribute-inference AUC (Fawcett, 2006; Hand & Till, 2001) and normalized MI (Duan et al., 2024; Noorbakhsh et al., 2024)). Experiments cover multiple cloud LLMs and edge SLMs. More details about metrics and models are provided in Tables 5 6 7 of Appendices C.3, C.4.

Table 1: Overall performance under a fixed global privacy budget B_{pri} and failure probability δ.

Method Family	Method	P-PubMedQA					P-FinQA					P-MultiHopRAG				
		Answer Quality			Leakage		Answer Quality			Leakage		Answer Quality			Leakage	
		F1↑	SM↑	LaJ↑	AUC↓	MI↓	F1↑	SM↑	LaJ↑	AUC↓	MI↓	F1↑	SM↑	LaJ↑	AUC↓	MI↓
LLM1+SLM1																
Cloud-Centric	Cloud-Only	.667	.762	.745	-	-	.741	.745	.795	-	-	.604	.787	.743	-	-
	RemoteRAG	.619	.661	.687	.687	.187	.684	.703	.748	.676	.221	.534	.691	.688	.683	.225
Edge–Cloud Collaboration	MINIONS	.591	.593	.678	.627	.164	.647	.682	.711	.634	.207	.501	.624	.594	.616	.206
	CoGenesis	.532	.571	.637	.606	.157	.609	.658	.696	.598	.191	.519	.617	.591	.614	.198
	LSRP	.496	.562	.612	.572	.142	.593	.646	.695	.587	.183	.495	.613	.579	.589	.189
	PRISM	.552	.575	.639	.559	.134	.628	.669	.703	.561	.176	.538	.684	.646	.554	.185
Edge-Only	Edge-Only	.445	.547	.531	.500	.000	.541	.615	.655	.500	.000	.458	.586	.547	.500	.000
Step-Wise	**Swim** (*Ours*)	.635	.747	.723	.550	.082	.691	.729	.783	.523	.158	.601	.772	.725	.541	.126
LLM1+SLM2																
Cloud-Centric	Cloud-Only	.671	.692	.821	-	-	.659	.786	.794	-	-	.554	.751	.786	-	-
	RemoteRAG	.607	.649	.744	.679	.214	.585	.684	.722	.669	.194	.503	.684	.724	.649	.188
Edge–Cloud Collaboration	MINIONS	.581	.637	.703	.662	.196	.567	.639	.677	.658	.167	.482	.653	.673	.634	.169
	CoGenesis	.538	.573	.636	.627	.165	.524	.583	.635	.627	.152	.463	.611	.648	.621	.146
	LSRP	.511	.545	.608	.593	.149	.502	.546	.593	.598	.145	.451	.574	.614	.597	.137
	PRISM	.573	.624	.682	.582	.137	.582	.667	.702	.581	.133	.480	.658	.679	.572	.162
Edge-Only	Edge-Only	.488	.516	.574	.500	.000	.493	.524	.561	.500	.000	.442	.531	.563	.500	.000
Step-Wise	**Swim** (*Ours*)	.656	.669	.812	.578	.135	.634	.763	.761	.571	.124	.512	.729	.766	.539	.123

5.2 OVERALL PERFORMANCE COMPARISON

Table 1 compares answer quality and empirical leakage under the same global privacy budget B_{pri} and failure probability δ. Across all six dataset-model settings, Swim achieves the highest F1, SM, and LaJ among the privacy-preserving baselines, while attaining the lowest AUC and MI among methods that communicate with the cloud. Averaged over all settings, Swim improves F1, SM, and LaJ over *RemoteRAG*, the strongest non-*Cloud-Only* baseline in macro-averaged answer quality, by 3.3%, 5.6%, and 4.3% absolute points, respectively. It also reduces AUC and MI by 0.124 and 0.081 (18.3% and 39.1% relative reductions). Compared with *Cloud-Only*, Swim retains 95.7% of F1 and 97.6% of LaJ, respectively. Although *Edge-Only* has the smallest leakage, its consistently lower answer-quality scores demonstrate the cost of avoiding cloud collaboration altogether.

The advantage is stable across domains and edge models. The largest F1 gain over *RemoteRAG* occurs on P-MultiHopRAG under LLM1+SLM1 (0.601 versus 0.534; $+0.067$). The largest AUC

Table 2: Empirical privacy leakage from cumulative cloud-visible transcript prefixes.

Metric	Model	P-PubMedQA				P-FinQA				P-MultiHopRAG			
		Init.	DP Rel.	Meta.	Full	Init.	DP Rel.	Meta.	Full	Init.	DP Rel.	Meta.	Full
AUC↓	LLM1+SLM1	.543	.532	.547	.550	.508	.514	.518	.523	.532	.536	.539	.541
	LLM1+SLM2	.543	.550	.555	.578	.542	.545	.551	.571	.528	.530	.533	.539
	LLM2+SLM1	.592	.553	.567	.596	.607	.578	.532	.585	.633	.611	.556	.555
	LLM2+SLM2	.594	.563	.537	.541	.608	.597	.533	.561	.540	.540	.509	.504
MI↓	LLM1+SLM1	.116	.113	.097	.082	.112	.071	.141	.158	.108	.146	.129	.126
	LLM1+SLM2	.111	.113	.126	.135	.129	.129	.126	.124	.042	.075	.062	.053
	LLM2+SLM1	.099	.090	.068	.062	.149	.141	.106	.097	.104	.084	.064	.071
	LLM2+SLM2	.095	.095	.068	.068	.129	.125	.089	.092	.113	.108	.080	.078

reduction relative to the lowest-AUC non-edge baseline occurs on P-FinQA under LLM1+SLM1 (0.523 for Swim versus 0.561 for PRISM), while the largest MI reduction occurs on P-MultiHopRAG under LLM1+SLM2 (0.053 for Swim versus 0.137 for LSRP). Replacing SLM1 with SLM2 changes Swim's macro F1 from 0.642 to 0.601 and macro AUC from 0.538 to 0.563; nevertheless, Swim remains the highest-quality privacy-preserving method and the lowest-leakage cloud-communicating method in all six settings. Results under additional LLM–SLM configurations are reported in Appendix D.

5.3 Privacy Leakage Analysis across Cloud-Visible Releases

Table 2 compares four cumulative transcript views: *Init.* contains sanitized planning exchanges; *DP Rel.* adds protected-release exchanges; *Meta.* adds public-exchange metadata, such as node/action IDs and payload structure; and *Full* includes their content. The last two are analysis views, not separate transmissions.

Mean AUC/MI estimates change from $0.564/0.109$ at *Init.* to $0.554/0.096$ at *Full*. These empirical fluctuations do not imply decreasing accounted privacy loss.

Model pairs produce different protected releases and cloud responses, changing the cues available to attackers. LLM2+SLM1 has the highest mean full-transcript AUC (0.579) but the lowest MI estimate (0.077); thus, no pair is uniformly most revealing across both measures under Swim's shared budget.

Domain differences reflect the attack target: P-FinQA has the highest mean MI estimate (0.118), as task-relevant financial terms can reveal its private topic, whereas P-MultiHopRAG has the lowest mean AUC (0.535), consistent with answers requiring evidence across multiple steps. Further analysis appears in Appendix E.

5.4 Ablation Study

We evaluate controlled Swim ablations of node-state quality estimation and future-aware privacy reservation. Unless explicitly ablated, all variants use the same data split, query graph, LLM–SLM pair, random seed, global privacy budget B_{pri}, and failure probability δ.

Node-state quality estimation. Starting from the full analytical estimator, *Heuristic Quality* replaces action-specific distortion estimates with fixed action-type heuristics and does not use U_i, I_i, or D_i. The *Fixed* U_i, *Fixed* I_i, and *Fixed* D_i variants replace exactly one node-dependent signal by the prespecified constant 0.5, while retaining the remaining components. These variants test whether each signal provides information beyond a dataset-level constant.

Future-reservation ablation. *No Future Reservation* sets future-node importance to zero, removing budget reservation for unresolved descendants.

Figure 3 shows that Full Swim consistently outperforms all ablations. Heuristic action scores cannot adapt to node-specific reasoning states, while the fixed-signal variants confirm that uncertainty, task importance, and private-information dependency provide complementary evidence. The larger impact of fixing I_i suggests that identifying globally consequential nodes is critical for directing limited privacy budget. Removing future reservation causes the strongest degradation because locally attractive actions may exhaust the budget before important downstream nodes become ready.

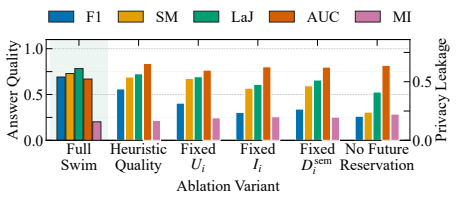

Figure 3: **Ablation study.** All ablations reduce answer quality and increase empirical leakage relative to Full Swim.

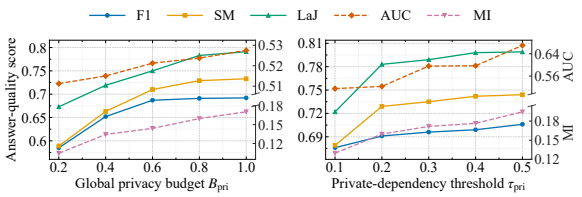

Figure 4: **Sensitivity analysis.** Quality saturates near $B_{\text{pri}} = 0.8$, while leakage rises sharply beyond $\tau_{\text{pri}} = 0.2$.

5.5 Sensitivity Analysis

Sensitivity to the global privacy budget. Increasing B_{pri} enlarges the set of feasible protected-cloud actions and initially improves answer quality. Beyond $B_{\text{pri}} = 0.8$, however, the most valuable reasoning steps have already received cloud assistance, so additional budget is mainly assigned to actions with lower marginal gain while leakage continues to increase. This diminishing-return pattern motivates selecting $B_{\text{pri}} = 0.8$.

Sensitivity to the privacy-dependency threshold. Increasing τ_{pri} allows more weakly privacy-dependent nodes to use privacy-free cloud execution, producing modest quality gains but greater leakage. In particular, increasing τ_{pri} from 0.2 to 0.3 raises AUC by 7.3% for less than 0.6% quality improvement, suggesting that some mildly privacy-dependent nodes no longer receive sufficient protection. We therefore select $\tau_{\text{pri}} = 0.2$.

5.6 Overhead Analysis

Table 3: Latency and cloud token costs under LLM1+SLM1. Parenthesized values show percentage changes relative to Cloud-Only; Lower is better.

Method	P-PubMedQA			P-FinQA			P-MultiHopRAG		
	Lat.(s)	Cloud In.	Cloud Out.	Lat.(s)	Cloud In.	Cloud Out.	Lat.(s)	Cloud In.	Cloud Out.
Cloud-Only	52.67	231	4934	86.54	136	8587	78.51	265	7024
RemoteRAG	58.71 (+11.5%)	328 (+42.0%)	5354 (+8.5%)	93.22 (+7.7%)	264 (+94.1%)	8799 (+2.5%)	86.56 (+10.3%)	289 (+9.1%)	7130 (+1.5%)
MINIONS	69.32 (+31.6%)	251 (+8.7%)	6194 (+25.5%)	110.39 (+27.6%)	156 (+14.7%)	9965 (+16.0%)	118.65 (+51.1%)	274 (+3.4%)	9259 (+31.8%)
CoGenesis	64.36 (+22.2%)	276 (+19.5%)	5854 (+18.6%)	99.64 (+15.1%)	172 (+26.5%)	9539 (+11.1%)	106.68 (+35.9%)	287 (+8.3%)	8827 (+25.7%)
LSRP	62.25 (+18.2%)	249 (+7.8%)	5459 (+10.6%)	98.47 (+13.8%)	148 (+8.8%)	9117 (+6.2%)	101.29 (+29.0%)	269 (+1.5%)	8301 (+18.2%)
PRISM	59.56 (+13.1%)	244 (+5.6%)	5968 (+21.0%)	100.69 (+16.4%)	179 (+31.6%)	9728 (+13.3%)	109.37 (+39.3%)	283 (+6.8%)	8948 (+27.4%)
Swim (*Ours*)	57.01 (+8.2%)	259 (+12.1%)	6255 (+26.8%)	94.61 (+9.3%)	192 (+41.2%)	9391 (+9.4%)	88.91 (+13.2%)	255 (-3.8%)	9154 (+30.3%)

Table 3 shows that Swim introduces moderate latency and cloud communication overhead relative to Cloud-Only, while remaining faster than MINIONS across all datasets. The additional cost mainly arises from local stochastic sampling, node-state estimation, and DP release preparation. Nevertheless, candidate actions are evaluated analytically and only the selected action is executed, avoiding repeated cloud trials and keeping the overall overhead bounded.

Finally, Appendix F presents a node-level case study of Swim's adaptive allocation process, while Appendix G demonstrates its deployment on a real-world mobile edge device.

6 Conclusion

We proposed Swim, a privacy- and executability-aware subquery routing framework for edge-cloud collaborative reasoning. Swim decomposes complex queries, estimates privacy dependency and edge executability, and combines hard privacy-first routing with soft score-based trade-offs. Experiments on three privacy-augmented QA datasets show that Swim improves answer quality over privacy-preserving baselines while maintaining low transmission-stage leakage. Ablation, overhead, and sensitivity analyses further confirm its effectiveness and robustness. We discuss the limitations and future work in Appendix H.

AI USE STATEMENT

In this work, we used generative AI tools to [assist with translation, assist in the writing of proofs]. We have not used generative AI tools to [generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, support qualitative and thematic data analysis, or interpret results], and [clean and reformat datasets] is designated as not applicable to this work. Additionally, we used generative AI tools for [sourcing/searching for information, identifying relevant literature, and proposing a title or keywords for the research paper]. We have reviewed all AI-assisted work. Specifically, we checked translations against the source text. For proofs drafted with AI assistance, the authors independently checked each inference against the stated assumptions, verified the intermediate mathematical derivations, and confirmed that the conclusions followed from those steps. We also verified suggested information and literature against original sources and reviewed the proposed title or keywords before using them. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses publicly available datasets and pretrained models.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, Sections 3–5 specify Swim's design, privacy-allocation procedure, and experimental protocol. Appendix C provides the dataset-construction and validation procedure, baseline definitions, cloud- and edge-model configurations, hardware settings, evaluation procedures, and privacy-accounting settings. The anonymous repository submitted as supplementary material contains the implementation, configuration files, data-processing pipeline, and evaluation scripts required to reproduce the reported experiments and analyses.

REFERENCES

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 17682–17690, 2024.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan R Routledge, et al. Finqa: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3697–3711, 2021.

Yihang Cheng, Lan Zhang, Junyang Wang, Mu Yuan, and Yunhao Yao. Remoterag: A privacy-preserving llm cloud rag service. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 3820–3837, 2025.

Chanyeol Choi, Jihoon Kwon, Jaeseon Ha, Hojun Choi, Chaewoon Kim, Yongjae Lee, Jy-yong Sohn, and Alejandro Lopez-Lira. Finder: Financial dataset for question answering and evaluating retrieval-augmented generation. In *Proceedings of the 6th ACM International Conference on AI in Finance*, pp. 638–646, 2025.

Lin Duan, Jingwei Sun, Jinyuan Jia, Yiran Chen, and Maria Gorlatova. Reimagining mutual information for enhanced defense against data leakage in collaborative inference. *Advances in Neural Information Processing Systems*, 37:44479–44500, 2024.

Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. doi: 10.1016/j.patrec.2005.10.010.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, Haofen Wang, et al. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1):32, 2023.

David J. Hand and Robert J. Till. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*, 45(2):171–186, 2001. doi: 10.1023/A:1010920819831.

Zixu Hao, Huiqiang Jiang, Shiqi Jiang, Ju Ren, and Ting Cao. Hybrid slm and llm for edge-cloud collaborative inference. In *Proceedings of the Workshop on Edge and Mobile Foundation Models*, pp. 36–41, 2024.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 2567–2577, 2019.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. Decomposed prompting: A modular approach for solving complex tasks. *arXiv preprint arXiv:2210.02406*, 2022.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. volume 33, pp. 9459–9474, 2020.

Senyao Li, Haozhao Wang, Wenchao Xu, Rui Zhang, Song Guo, Jingling Yuan, Xian Zhong, Tianwei Zhang, and Ruixuan Li. Collaborative inference and learning between edge slms and cloud llms: A survey of algorithms, execution, and open challenges. *arXiv preprint arXiv:2507.16731*, 2025.

Avanika Narayan, Sabri Eyuboglu, Dan Biderman, Avner May, Scott Linderman, James Zou, and Christopher Re. Cost-efficient collaboration between on-device and cloud language models. In *ICLR 2025 Workshop on Foundation Models in the Wild*, 2025.

Sayedeh Leila Noorbakhsh, Binghui Zhang, Yuan Hong, and Binghui Wang. {Inf2Guard}: An {Information-Theoretic} framework for learning {Privacy-Preserving} representations against inference attacks. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 2405–2422, 2024.

Sarah Packowski, Inge Halilovic, Jenifer Schlotfeldt, and Trish Smith. Optimizing and evaluating enterprise retrieval-augmented generation (rag): A content design perspective. In *Proceedings of the 2024 8th International Conference on Advances in Artificial Intelligence*, pp. 162–167, 2024.

Qi Pang, Jinhao Zhu, Helen Möllering, Wenting Zheng, and Thomas Schneider. Bolt: Privacy-preserving, accurate and efficient inference for transformers. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 4753–4771. IEEE, 2024.

Guanqiao Qu, Qiyuan Chen, Wei Wei, Zheng Lin, Xianhao Chen, and Kaibin Huang. Mobile edge intelligence for large language models: A contemporary survey. *IEEE Communications Surveys & Tutorials*, 27(6):3820–3860, 2025.

Yixuan Tang and Yi Yang. Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries. *arXiv preprint arXiv:2401.15391*, 2024.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 10014–10037, Toronto, Canada, 2023. Association for Computational Linguistics.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6233–6251, Bangkok, Thailand, 2024. Association for Computational Linguistics.

Justin Xu, Yiming Li, Zizheng Zhang, Augustine Yui Hei Luk, Mayank Jobanputra, Samarth Oza, Ashley Murray, Meghana Reddy Kasula, Andrew Parker, and David W Eyre. Tree-of-quote prompting improves factuality and attribution in multi-hop and medical reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5605–5622, 2025.

Rongjie Yi, Liwei Guo, Shiyun Wei, Ao Zhou, Shangguang Wang, and Mengwei Xu. Edgemoe: Empowering sparse large language models on mobile devices. *IEEE Transactions on Mobile Computing*, 24(8):7059–7073, 2025.

Ziqian Zeng, Jianwei Wang, Junyao Yang, Zhengdong Lu, Haoran Li, Huiping Zhuang, and Cen Chen. Privacyrestore: Privacy-preserving inference in large language models via privacy removal and restoration. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10821–10855, 2025.

Junfei Zhan, Haoxun Shen, Zheng Lin, and Tengjiao He. Prism: Privacy-aware routing for adaptive cloud–edge llm inference via semantic sketch collaboration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 28150–28158, 2026.

Kaiyan Zhang, Jianyu Wang, Ermo Hua, Biqing Qi, Ning Ding, and Bowen Zhou. Cogenesis: A framework collaborating large and small language models for secure context-aware instruction following. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4295–4312, 2024a.

Mingjin Zhang, Xiaoming Shen, Jiannong Cao, Zeyang Cui, and Shan Jiang. Edgeshard: Efficient llm inference via collaborative edge computing. *IEEE Internet of Things Journal*, 12(10):13119–13131, 2024b.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations (ICLR)*, 2020.

Yingyi Zhang, Pengyue Jia, Xianneng Li, Derong Xu, Maolin Wang, Yichao Wang, Zhaocheng Du, Huifeng Guo, Yong Liu, Ruiming Tang, and et al. LSRP: A leader-subordinate retrieval framework for privacy-preserving cloud-device collaboration. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 3889–3900. Association for Computing Machinery, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

A Related Work

Edge–Cloud Collaborative Generation. Edge–cloud collaboration combines the reasoning capability of cloud LLMs with the privacy, locality, and efficiency advantages of edge SLMs. CoGenesis enables secure context-aware generation through sketch- and logit-level collaboration between large and small models (Zhang et al., 2024a). LSRP adopts a leader–subordinate retrieval framework in which the cloud provides task-level guidance while private data remain on the device (Zhang et al., 2025). PRISM further performs sensitivity-aware routing and semantic-sketch collaboration between cloud and edge models (Zhan et al., 2026). Other systems primarily target efficiency: MINIONS decomposes data-intensive tasks for cost-efficient local execution (Narayan et al., 2025), Hybrid SLM–LLM selectively invokes cloud assistance during inference (Hao et al., 2024), and EdgeShard partitions model computation across heterogeneous edge and cloud resources (Zhang et al., 2024b). These methods establish the benefits of edge–cloud collaboration, but do not allocate a shared end-to-end privacy budget across interdependent reasoning steps. In contrast, Swim constructs privacy-valid execution actions and allocates privacy according to their state-dependent quality gains and incremental accountant costs.

Privacy-Preserving RAG and Private Inference. Privacy-preserving RAG and private inference aim to exploit remote models without directly exposing sensitive queries, documents, or intermediate representations. RemoteRAG protects cloud-based retrieval through perturbed representations and restricted retrieval (Cheng et al., 2025). BOLT provides privacy-preserving transformer inference through secure computation (Pang et al., 2024), while PrivacyRestore removes private information before remote inference and restores it locally afterward (Zeng et al., 2025). These approaches protect particular inputs or inference stages, but do not determine how limited privacy should be spent across a multi-step reasoning process. Swim instead keeps private knowledge at the trusted edge, protects private-dependent releases with differential privacy, composes their costs through an online accountant, and allocates a global privacy budget according to the predicted return of each reasoning action.

Query Decomposition and Multi-Step Reasoning. Complex reasoning is commonly organized into multiple dependent steps. Decomposed Prompting delegates simpler subproblems to specialized prompts, models, or external modules (Khot et al., 2022). IRCoT interleaves retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions (Trivedi et al., 2023), while Graph of Thoughts represents intermediate reasoning states as a graph rather than a linear chain (Besta et al., 2024). These studies focus primarily on constructing or executing structured reasoning processes.

B Additional Theoretical Analysis

This appendix provides additional details for the theoretical results in Section 3. We first clarify the conditional interpretation and monotonicity of operational privacy dependency, then prove the privacy–utility lower bound, and finally connect adaptive privacy composition with the online allocation rule in Section 4.4.

B.1 Conditional Interpretation and Monotonicity

The trusted edge first releases a sanitized query $\widetilde{Q}$, from which the cloud constructs the reasoning DAG $G = \mathsf{Plan}(\widetilde{Q}, K_{\mathrm{pub}})$. Because G is computed only from already cloud-visible information and public knowledge, its construction is post-processing and incurs no additional privacy cost beyond the release of $\widetilde{Q}$. If the sanitization mechanism depends on protected information, its privacy cost is included as an initial release in the end-to-end accountant.

After graph construction, the node-level analysis conditions on the realized DAG G and the cloud-visible context X_i available before executing node v_i. The context X_i contains $\widetilde{Q}$, G, and all preceding cloud-visible releases. For any fixed $X_i = x_i$, the node protocol Γ_i must satisfy its stated (ϵ_i, δ_i)-DP guarantee with respect to adjacent private knowledge bases in $\mathcal{K}_i(x_i)$. Thus, ϵ_i represents the incremental privacy cost of the current protocol, whereas privacy already incurred in constructing X_i is handled by end-to-end accounting.

For completeness, operational privacy dependency is monotone in the allowed distortion. Let

$$\mathfrak{G}_i(D) = \left\{ \Gamma_i : \sup_{K_{\mathrm{pri}} \in \mathcal{K}_i(x_i)} \mathbb{E}\left[d_i\Big(Y_i(K_{\mathrm{pri}}), \widehat{Y}_i^c\Big)\right] \le D \right\}$$

denote the protocols achieving distortion at most D. If $D_1 \le D_2$, then

$$\mathfrak{G}_i(D_1) \subseteq \mathfrak{G}_i(D_2).$$

Taking the infimum privacy cost over these nested feasible sets gives

$$\rho_i(D_1; \delta_i) \ge \rho_i(D_2; \delta_i).$$

Hence, allowing larger distortion cannot increase the minimum privacy budget, and $\rho_i(D; \delta_i)$ characterizes a nonincreasing privacy–utility frontier.

B.2 Proof of the Privacy–Utility Lower Bound

We prove Theorem 2. Throughout the proof, we condition on a fixed cloud-visible context $X_i = x_i$ and omit it from the notation.

Consider any adjacent pair

$$K_{\mathrm{pri}} \simeq K'_{\mathrm{pri}}$$

in $\mathcal{K}_i(x_i)$, with target answers

$$y = Y_i(K_{\mathrm{pri}}), \qquad y' = Y_i(K'_{\mathrm{pri}}),$$

and let

$$\Delta = d_i(y, y').$$

Let P and Q denote the distributions of the cloud-assisted output $\widehat{Y}_i^c$ under K_{pri} and K'_{pri}, respectively. By post-processing and the symmetry of the adjacency relation, (ϵ_i, δ_i)-DP implies that, for every measurable event A,

$$P(A) \le e^{\epsilon_i} Q(A) + \delta_i, \qquad Q(A) \le e^{\epsilon_i} P(A) + \delta_i. \tag{11}$$

We first bound the total-variation distance between P and Q.

Lemma 1 (DP-Induced Total-Variation Bound)**.** *The distributions P and Q satisfy*

$$\mathrm{TV}(P, Q) \le \frac{e^{\epsilon_i} - 1 + 2\delta_i}{e^{\epsilon_i} + 1}. \tag{12}$$

Proof. Let $a = e^{\epsilon_i}$. For an arbitrary event A, assume without loss of generality that

$$p = P(A) \ge q = Q(A),$$

and define $t = p - q$. From Eq. (11),

$$p \le aq + \delta_i,$$

which gives

$$(a-1)q \ge t - \delta_i. \tag{13}$$

Applying the reverse DP inequality to the complementary event A^c gives

$$1 - q \le a(1-p) + \delta_i,$$

or equivalently,

$$(a-1)q + at \le a - 1 + \delta_i. \tag{14}$$

Combining Eqs. (13) and (14) yields

$$(a+1)t \le a - 1 + 2\delta_i.$$

Taking the supremum over measurable events A proves Eq. (12). □

By the maximal coupling theorem, there exist random variables

$$Z \sim P, \qquad Z' \sim Q$$

such that

$$\Pr[Z = Z'] = 1 - \mathrm{TV}(P, Q).$$

On the event $Z = Z'$, let their common value be z. Since d_i satisfies the triangle inequality,

$$d_i(y, z) + d_i(y', z) \geq d_i(y, y') = \Delta.$$

Therefore,

$$\mathbb{E}_P \left[d_i(y, \widehat{Y}_i^c) \right] + \mathbb{E}_Q \left[d_i(y', \widehat{Y}_i^c) \right] \geq \Delta \left(1 - \mathrm{TV}(P, Q) \right). \tag{15}$$

By the uniform distortion assumption in Theorem 2,

$$\mathbb{E}_P \left[d_i(y, \widehat{Y}_i^c) \right] \leq D_i, \qquad \mathbb{E}_Q \left[d_i(y', \widehat{Y}_i^c) \right] \leq D_i.$$

Consequently,

$$2D_i \geq \Delta \left(1 - \mathrm{TV}(P, Q) \right).$$

Using Lemma 1,

$$1 - \mathrm{TV}(P, Q) \geq \frac{2(1 - \delta_i)}{1 + e^{\epsilon_i}},$$

and hence

$$D_i \geq \frac{(1 - \delta_i)\Delta}{1 + e^{\epsilon_i}}.$$

Because this inequality holds for every adjacent pair, taking the supremum over $K_{\mathrm{pri}} \simeq K'_{\mathrm{pri}}$ gives

$$D_i \geq \frac{(1 - \delta_i)\Delta_i}{1 + e^{\epsilon_i}},$$

which proves Theorem 2.

Necessary privacy budget. For $D_i > 0$, rearranging the bound gives

$$e^{\epsilon_i} \geq \frac{(1 - \delta_i)\Delta_i}{D_i} - 1.$$

Therefore, whenever

$$D_i < \frac{(1 - \delta_i)\Delta_i}{2},$$

every protocol attaining distortion at most D_i must satisfy

$$\epsilon_i \geq \log \left(\frac{(1 - \delta_i)\Delta_i}{D_i} - 1 \right). \tag{16}$$

Taking the infimum over all feasible protocols yields

$$\rho_i(D_i; \delta_i) \geq \left[\log \left(\frac{(1 - \delta_i)\Delta_i}{D_i} - 1 \right) \right]_+,$$

where $[x]_+ = \max\{x, 0\}$. This expression gives a necessary, rather than sufficient, privacy budget for achieving the target distortion.

B.3 Adaptive Composition and Allocation Derivation

Let

$$M_{1:T}^{e\to c} = (M_1^{e\to c}, \dots, M_T^{e\to c})$$

denote the complete sequence of cloud-visible edge-to-cloud releases. When the initial sanitized query $\widetilde{Q}$ depends on protected information, it is included as one of these releases. The t-th release may depend adaptively on the preceding interaction history H_{t-1}.

For every fixed history, suppose that

$$\Pr\left[M_t^{e\to c} \in A \mid H_{t-1}, K_{\mathrm{pri}}\right] \le e^{\epsilon_t} \Pr\left[M_t^{e\to c} \in A \mid H_{t-1}, K'_{\mathrm{pri}}\right] + \delta_t$$

for every measurable event A and every adjacent pair $K_{\mathrm{pri}} \simeq K'_{\mathrm{pri}}$. Adaptive sequential composition then gives

$$M_{1:T}^{e\to c} \text{ is } \left(\sum_{t=1}^{T} \epsilon_t, \sum_{t=1}^{T} \delta_t\right)\text{-DP.} \tag{17}$$

Independence among releases is not required. Furthermore, cloud-side computation based only on previously released information is post-processing and incurs no additional privacy loss. However, routing decisions, invocation patterns, termination behavior, or other metadata must be included in the cloud-visible transcript whenever their distributions depend on protected information.

A node action may contain multiple atomic releases. The global accountant therefore operates over the atomic release history and reports

$$\mathcal{P}(\pi) = \epsilon_{\mathrm{acct}}(\pi; \delta)$$

for the prescribed global failure probability δ. Let H_i be the accountant state before executing ready node v_i. The incremental cost of candidate action a is

$$p_i(a) = \epsilon_{\mathrm{acct}}\left(H_i \oplus a; \delta\right) - \epsilon_{\mathrm{acct}}\left(H_i; \delta\right),$$

where $H_i \oplus a$ denotes the hypothetical state after adding every atomic release associated with a.

To connect the global constrained problem with node-level allocation, consider the change in the Lagrangian caused by executing candidate action a at the current node. Under the local analytic approximation, its predicted utility increment is $\widehat{g}_i(a)$ and its incremental privacy cost is $p_i(a)$. The corresponding one-step Lagrangian increment is therefore

$$\Delta\widehat{\mathcal{L}}_i(a; \lambda) = \widehat{g}_i(a) - \lambda p_i(a). \tag{18}$$

If the future node states were known, λ would represent the marginal value of the remaining global privacy budget. In an online reasoning DAG, however, the uncertainty and private-information dependency of unresolved nodes are unavailable before their predecessor outputs are realized. Swim therefore instantiates the marginal price using the static importance of unresolved descendants.

Specifically, let

$$B_i^{\mathrm{rem}} = B_{\mathrm{pri}} - \epsilon_{\mathrm{acct}}(H_i; \delta)$$

be the remaining privacy margin and let

$$I_i^{\mathrm{future}} = \sum_{v_j \in \mathcal{V}_i^{\mathrm{unresolved}}} \frac{S_j}{1 + d(i,j)}$$

be the aggregate discounted importance of unresolved descendants. The state-dependent privacy price is

$$\lambda_i^{\star} = \frac{I_i^{\mathrm{future}}}{\max\left\{B_i^{\mathrm{rem}}, \eta_B\right\}}.$$

This quantity is a receding-horizon surrogate for the exact dual price rather than a claim of globally optimal dual recovery. It becomes larger when more important downstream reasoning remains or when the available privacy margin shrinks, thereby discouraging low-return privacy expenditure at the current node.

Define the accountant-feasible action set

$$\mathcal{F}_i = \{a \in \mathcal{A}_i^\star : \epsilon_{\text{acct}}(H_i \oplus a; \delta) \le B_{\text{pri}}\}.$$

The online action is then selected as

$$a_i^\star = \arg\max_{a \in \mathcal{F}_i} [\widehat{g}_i(a) - \lambda_i^\star p_i(a)]. \tag{19}$$

Only the selected action $a_i^\star$ is executed. Candidate gains are estimated analytically from local statistics, so evaluating unselected actions creates neither cloud calls nor additional privacy loss. After execution, the accountant state and reasoning context are updated, newly ready nodes are evaluated, and the future reserve and privacy price are recomputed. The hard feasibility constraint guarantees that errors in the analytic quality model cannot cause the cumulative accounted privacy loss to exceed B_{pri}.

C Experimental Details

C.1 Dataset Construction

We construct *P-PubMedQA*, *P-FinQA*, and *P-MultiHopRAG* from PubMedQA Jin et al. (2019), FinQA Chen et al. (2021), and MultiHopRAG Tang & Yang (2024), respectively. We follow an evidence-grounded augmentation procedure that incorporates private information into the reasoning chain and jointly revises the question and reference answer.

Evidence identification. For each original instance, annotators first identify the evidence passages and reasoning chain supporting its answer. These establish the entities, relations, and events to which the additional private information must be connected.

Private evidence construction. Annotators design a domain-consistent private attribute linked to the same entity, relation, or event and introduce it as a necessary evidence step. Examples include a patient-specific clinical state in P-PubMedQA, an internal financial indicator in P-FinQA, and a non-public event relation in P-MultiHopRAG. Public evidence forms $\mathcal{K}_{pub}$, while the constructed private evidence is assigned to the edge-side $\mathcal{K}_{pri}$.

Joint question and answer revision. The question and reference answer are rewritten together to incorporate the added evidence dependency. This ensures that the private attribute contributes to the reasoning required by the question, rather than appearing as an unrelated field in an otherwise unchanged instance.

Private-evidence necessity check. An instance is retained only when the combined public and private evidence supports its reference answer, while removing the private evidence leaves the answer incomplete or indeterminate. This criterion establishes a substantive dependency on private knowledge while preserving the source task's reasoning structure.

Human Validation Domain-knowledgeable annotators verify evidence consistency and answer dependency. They check whether the added private attribute is consistent with the source evidence, whether the rewritten question and reference answer agree with the combined evidence, and whether the private evidence satisfies the necessity criterion above. Disagreements are resolved by majority vote.

C.2 Baseline Details

We compare Swim with cloud-only, edge-only, and representative edge–cloud collaboration methods, as summarized in Table 4.

Cloud-Only represents the unconstrained-quality reference, whereas Edge-Only represents the no-communication reference. The remaining baselines cover query protection, selective transmission, representation-level collaboration, cloud guidance, and sensitivity-aware routing.

Table 4: Baselines spanning cloud-only, edge–cloud collaborative, and edge-only execution.

Type	Baseline	Description
Cloud-Only	Cloud-Only	Executes every reasoning step on the cloud with full access to private inputs, providing a privacy-unconstrained quality upper bound.
	RemoteRAG (Cheng et al., 2025)	Protects user queries through perturbed embeddings and restricted cloud retrieval.
Edge-Cloud	MINIONS (Narayan et al., 2025)	Selectively transmits local context to a cloud LLM for data-intensive reasoning.
	CoGenesis (Zhang et al., 2024a)	Uses cloud-generated sketches or logits to guide edge-side generation.
	LSRP (Zhang et al., 2025)	Uses privacy-preserving cloud guidance while retaining private data on the edge.
	PRISM (Zhan et al., 2026)	Combines sensitivity-aware routing, local differential privacy, and edge-side refinement.
Edge-Only	Edge-Only	Executes every reasoning step using the edge SLM without cloud communication.

C.3 Implementation Details

We evaluate Swim using three cloud-side LLMs and three edge-side SLMs, covering both open-source and proprietary cloud models as well as desktop- and mobile-class edge deployments. Table 5 summarizes the model configurations.

Table 5: Cloud-side LLM and edge-side SLM configurations.

Cloud-Side LLMs			Edge-Side SLMs		
No.	Model	Source	No.	Model	Deployment
LLM1	Qwen3.5-397B	Open-Source	SLM1	ChatGLM-6B	Desktop
LLM2	GLM-5.2	Closed-Source	SLM2	Qwen2.5-7B	Desktop
LLM3	GPT-5.4	Closed-Source	SLM3	GLM-Edge-1.5B	Mobile

LLM1–LLM3 represent cloud models with different architectures, access modes, and reasoning capabilities. SLM1 and SLM2 are deployed on desktop GPUs to evaluate Swim under different local model capacities, whereas SLM3 is a lightweight model used to assess deployment on resource-constrained mobile devices. Each LLM–SLM pair uses the same datasets, privacy configurations, and evaluation protocol to ensure comparability.

Table 6: Principal implementation parameters used in the reported experiments.

Component	Setting
Local stochastic executions	$M = 5$ executions per ready node
Hidden-state pooling	$K = 4$ adaptive pooling segments per structured interval
Stochastic decoding	Temperature 0.8, top-p 0.95, and at most 128 generated tokens for node-state estimation
Private-dependency estimation	Gaussian-kernel bandwidth $h = 1.0$ and dependency threshold $\tau_{\mathrm{pri}} = 0.2$
Global privacy constraint	$B_{\mathrm{pri}} = 0.8$ and target failure probability $\delta = 10^{-4}$
Protected-release configurations	$\epsilon_\ell \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$ with per-release $\delta_\ell = 10^{-6}$
Protected state release	State dimension 64, clipping norm $C = 1.0$, and at most 16 release units per node
Privacy attack	Five out-of-fold splits, random seed 6, and a 512-dimensional character-n-gram representation

The selected values of B_{pri} and τ_{pri} follow the sensitivity analysis in Section 5.5. They are fixed across the main comparisons and are not adjusted for individual records or nodes.

Desktop experiments use Python 3.10, PyTorch 2.1, CUDA 12.1, Hugging Face Transformers 4.40, and Accelerate. Local inference uses FP16 and FlashAttention when supported. Runtime dependencies are packaged in NVIDIA CUDA runtime containers, while the host GPU driver is exposed through the NVIDIA container runtime and model checkpoints are mounted separately. Cloud models are accessed through HTTPS-based JSON APIs.

Hardware-specific latency is measured on NVIDIA RTX 3080, RTX 2080 Ti, and RTX 2080 GPUs. Real-device deployment are reported separately in G; the corresponding runtime backend should be named only after the final mobile measurements are verified.

C.4 METRICS AND PRIVACY EVALUATION

Table 7 summarizes the answer-quality, empirical-leakage, and formal-accounting metrics. F1, SoftMatch, and LLM-as-a-Judge evaluate complementary aspects of answer correctness, whereas AUC and MI measure information recoverable by an empirical attacker.

Table 7: Evaluation metrics for answer quality, empirical privacy leakage, and formal privacy accounting.

Type	Metric	Description
Answer Quality	F1 (Zhang et al., 2020)	Measures contextual semantic similarity between the prediction and reference using BERTScore.
	SoftMatch (SM) (Xu et al., 2025)	Measures answer-level semantic equivalence between the prediction and reference.
	LLM-as-a-Judge (LaJ) (Zheng et al., 2023)	Correctness and relevance assessed by a fixed strong LLM judge.
Empirical Leakage	Attribute-Inference AUC (Fawcett, 2006; Hand & Till, 2001)	Macro one-vs-rest ROC-AUC of a private-label attacker evaluated on held-out cloud-visible transcripts.
	Mutual Information (MI) (Duan et al., 2024; Noorbakhsh et al., 2024)	Normalized variational lower-bound estimate of the private-label information recoverable from cloud-visible transcripts.

D OVERALL PERFORMANCE ON OTHER LLM+SLM SETTINGS

Table 8 extends the fixed-budget comparison to LLM2/LLM3 and SLM1/SLM2/SLM3. The table now reports answer quality, empirical leakage, and accounted privacy loss under the same B_{pri} and δ.

Table 8: Additional performance under a fixed global privacy budget B_{pri} and failure probability δ.

Method Family	Method	P-PubMedQA					P-FinQA					P-MultiHopRAG				
		Answer Quality			Leakage		Answer Quality			Leakage		Answer Quality			Leakage	
		F1↑	SM↑	LaJ↑	AUC↓	MI↓	F1↑	SM↑	LaJ↑	AUC↓	MI↓	F1↑	SM↑	LaJ↑	AUC↓	MI↓
LLM2+SLM1																
Cloud-Centric	Cloud-Only	.698	.733	.801	-	-	.760	.792	.797	-	-	.584	.809	.815	-	-
	RemoteRAG	.612	.649	.714	.662	.195	.675	.692	.713	.647	.228	.512	.703	.757	.652	.217
Edge–Cloud Collaboration	MINIONS	.581	.617	.671	.634	.181	.646	.654	.694	.623	.217	.498	.687	.721	.631	.205
	CoGenesis	.559	.598	.637	.594	.168	.589	.611	.637	.593	.203	.486	.622	.658	.603	.191
	LSRP	.508	.551	.617	.558	.148	.526	.584	.621	.568	.194	.471	.605	.609	.586	.178
	PRISM	.588	.629	.708	.547	.135	.648	.652	.698	.557	.172	.508	.692	.739	.550	.161
Edge-Only	Edge-Only	.465	.534	.594	.500	.000	.491	.552	.597	.500	.000	.442	.589	.637	.500	.000
Step-Wise	**Swim** (*Ours*)	.651	.705	.794	.538	.112	.733	.761	.755	.542	.137	.548	.786	.788	.541	.122
LLM2+SLM2																
Cloud-Centric	Cloud-Only	.647	.687	.825	-	-	.711	.782	.806	-	-	.543	.752	.793	-	-
	RemoteRAG	.593	.642	.773	.623	.178	.642	.717	.736	.652	.197	.501	.692	.716	.643	.218
Edge–Cloud Collaboration	MINIONS	.576	.631	.742	.617	.163	.607	.689	.703	.637	.186	.485	.665	.695	.611	.209
	CoGenesis	.552	.614	.713	.604	.149	.585	.675	.682	.623	.159	.469	.624	.683	.597	.196
	LSRP	.529	.596	.672	.586	.132	.556	.628	.651	.588	.142	.452	.582	.637	.588	.186
	PRISM	.578	.622	.748	.581	.128	.649	.728	.739	.561	.131	.492	.683	.705	.562	.185
Edge-Only	Edge-Only	.513	.582	.638	.500	.000	.532	.578	.616	.500	.000	.433	.559	.619	.500	.000
Step-Wise	**Swim** (*Ours*)	.623	.658	.792	.558	.096	.682	.754	.783	.547	.109	.522	.719	.757	.549	.163
LLM3+SLM3																
Cloud-Centric	Cloud-Only	.694	.766	.731	1.0	1.0	.755	.779	.787	1.0	1.0	.623	.749	.721	1.0	1.0
	RemoteRAG	.615	.672	.696	.684	.193	.691	.694	.744	.676	.217	.545	.696	.684	.672	.216
Edge–Cloud Collaboration	MINIONS	.588	.602	.669	.621	.167	.642	.673	.713	.634	.201	.513	.631	.627	.611	.202
	CoGenesis	.541	.558	.644	.611	.153	.603	.661	.687	.598	.189	.501	.603	.582	.605	.192
	LSRP	.503	.542	.625	.575	.139	.591	.629	.681	.587	.182	.492	.588	.563	.593	.178
	PRISM	.604	.658	.679	.562	.132	.685	.688	.725	.566	.171	.537	.667	.654	.572	.175
Edge-Only	Edge-Only	.467	.524	.573	.500	.000	.534	.605	.562	.500	.000	.463	.569	.536	.500	.000
Step-Wise	**Swim** (*Ours*)	.662	.717	.712	.541	.121	.738	.756	.761	.543	.161	.594	.729	.732	.547	.168

E Further Analysis of Privacy Leakage

E.1 Privacy Leakage under Different Attacker Strengths

Table 9 compares attack models on the same full cloud-visible transcript and under the same fixed privacy budget.

Table 9: Privacy leakage under attackers of different strengths on the full cloud-visible transcript.

Strength	Attack Model	P-PubMedQA		P-FinQA		P-MultiHopRAG	
		AUC	MI	AUC	MI	AUC	MI
Weak	Bag-of-Words Attacker	.528	.069	.517	.113	.527	.106
Medium	Encoder-Based Attacker	.538	.074	.519	.124	.534	.111
Strong	LLM-Based Attacker	.550	.082	.523	.158	.541	.126

E.2 Content Analysis of Privacy Leakage

Table 10 reports leakage by sensitive-content category together with the number of evaluated samples. The previous AUC and MI values are retained as placeholders; category counts and all leakage values will be updated from the revised transcript-level evaluation.

Table 10: Privacy leakage by sensitive-content category.

Category	Dataset	AUC	MI
Cardiometabolic and Organ Function	P-PubMedQA	.588	.166
Private Person/Organization Identity	P-MultiHopRAG	.577	.158
Private Comparison/Temporal Judgment	P-MultiHopRAG	.560	.147
Behavioral and Functional Health	P-PubMedQA	.549	.145
Debt and Liquidity	P-FinQA	.544	.136

F Case Study

This case study illustrates the execution of Swim on a concrete metabolic-risk assessment query. Gray and blue boxes denote edge-side and cloud-side execution, respectively.

Privacy-Safe DAG Construction **[Edge → Cloud]**

`Original Query:`

Based on my local health record including the records of LDL cholesterol, fasting glucose, body weight, and waist circumference from my personal health examinations and dietary history, together with public guidelines on metabolic syndrome and cardiovascular risk, determine whether my current metabolic risk.

`Desensitized Query:`

The cloud decomposes the desensitized query into five dependent reasoning nodes without accessing the original health record.

Node 1: Personal Trend Retrieval **[Edge]**

`Query:`

Retrieve the longitudinal changes in LDL cholesterol, fasting glucose, body weight, waist circumference, diet, and exercise.

`Private Evidence:`

LDL increases from 3.1 to 4.4 mmol/L, fasting glucose from 5.4 to 6.3 mmol/L, body weight from 68.2 to 74.6 kg, and waist circumference from 82 to 91 cm. High-calorie meals increase from once to four times per week, while exercise decreases from four times to once per week.

`Output:`

The protected trend summary indicates consistent deterioration in both metabolic indicators and lifestyle factors.

Node 2: Public Guideline Retrieval **[Public Cloud]**

`Query:`

Retrieve public guidelines and standards for metabolic-syndrome diagnosis and cardiovascular-risk assessment.

`Output:`

The retrieved evidence includes diagnostic thresholds for waist circumference, fasting glucose, triglycerides, and HDL cholesterol, together with the PAR and SCORE2 cardiovascular-risk assessment frameworks.

Node 3: Indicator–Guideline Matching **[Protected Cloud]**

`Query:`

Compare the observed indicator trends with the risk patterns specified by the public clinical guidelines.

`Output:`

The increases in LDL cholesterol, fasting glucose, body weight, and waist circumference jointly move the user toward clinically relevant metabolic-risk ranges.

Node 4: Metabolic Risk Analysis **[Edge]**

`Query:`

Based on the matched indicators, identify the likely causes and determine whether the current metabolic risk.

`Output:`

Weight gain, more frequent high-calorie meals, and reduced exercise are likely contributors to the deteriorating indicators. The combined trend warrants timely attention and lifestyle intervention.

Node 5: Final Evidence Fusion **[Edge]**

`Input:`

The original private measurements, the protected trend summary, the public clinical criteria, and the cloud-side risk analysis.

`Final Answer:`

The concurrent increases in LDL cholesterol, fasting glucose, body weight, and waist circumference indicate a rising metabolic risk. More frequent high-calorie meals, weight gain, and reduced exercise are likely contributing factors. The current condition warrants attention, and the risk may continue to accumulate without timely lifestyle intervention.



① The file path of the private knowledge base on the Android system of OPPO Find X8.
② Private daily records retrieved from the local knowledge base.
③ A sample query for retrieving personal medical-visit information.
④ A sample query for generating personalized diagnostic evaluation suggestions.
⑤ The edge-side SLM configuration on OPPO Find X8.

Figure 5: A real-world mobile device demonstration on OPPO Find X8.

G Real-World Mobile Device Demonstration

We demonstrate the mobile workflow of Swim on an OPPO Find X8, with GLM-Edge-1.5B serving as the on-device SLM (Fig. 5 ⑤). The phone stores private health records locally (Fig. 5 ①), allowing relevant notes and measurements to be retrieved without uploading the raw file (Fig. 5 ②). Figures 5 ③ and ④ show a retrieval query and a representative health-risk response, respectively.

In the collaborative workflow, the edge sanitizes the user query for cloud-side DAG planning. For each ready reasoning node, Swim estimates local uncertainty, task importance, and private-information dependency, then selects an edge or cloud action subject to the remaining global privacy budget. When a protected cloud action is selected, the phone sends the sanitized node request together with a DP-protected release to the cloud LLM API. The returned response is incorporated into subsequent reasoning on the phone; the raw private records remain on the device.

This demonstration illustrates how local private-knowledge access and on-device inference can be combined with selective, privacy-accounted cloud assistance in a mobile setting.

H Limitations

Swim uses repeated local SLM executions to estimate uncertainty and static node importance to reserve privacy budget for unresolved steps. These choices avoid speculative cloud calls and keep allocation tractable, but introduce local computation and do not account for every change in downstream reasoning needs. Future work could reduce sampling overhead through adaptive stopping, update future-node importance as the DAG is executed, and assess Swim on a broader range of reasoning tasks and edge devices. These extensions leave the step-wise allocation framework and its privacy accounting unchanged.