SPEAR: PRIVACY- AND EXECUTABILITY-AWARE SUBQUERY ROUTING FOR EDGE-CLOUD COLLABORATIVE RAG

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-augmented generation (RAG) in financial, medical, and enterprise applications often depends on private local data, creating a *structural mismatch*: edge devices can access these data but have limited reasoning capabilities, while powerful cloud models cannot directly access them. Existing edge-cloud methods largely provide coarse-grained sketches or guidance, limiting their effectiveness on multi-step queries involving mixed public-private knowledge. We propose **SPEAR**, a subquery-level, privacy- and executability-aware routing method for edge-cloud collaborative RAG. SPEAR desensitizes the user query for cloud-side decomposition into a dependency-structured subquery graph, then jointly estimates each subquery's privacy dependency and edge executability to select its execution site. Subqueries are executed collaboratively according to graph dependencies, and their outputs are fused at the edge. Experiments across multiple privacy-sensitive domain datasets show that SPEAR consistently improves generation quality while limiting private knowledge exposure. Code is available at `https://anonymous.4open.science/r/SPEAR-5D6D`.

1 INTRODUCTION

Retrieval-augmented generation (RAG) is widely used in knowledge-intensive applications by improving LLM generation with external retrieval (Lewis et al., 2020; Gao et al., 2023; Fan et al., 2024; Zhu et al., 2025; Yu et al., 2025). In privacy-sensitive domains such as financial analysis, medical consultation, and enterprise assistance, RAG often depends on sensitive data that cannot leave local devices or organizational boundaries (Xiong et al., 2024; Amugongo et al., 2025; Choi et al., 2025; Zhao et al., 2025; Packowski et al., 2024; Xu et al., 2024). Meanwhile, edge-side devices accessing such data typically lack the computational budget for high-quality LLM reasoning and generation (Qu et al., 2025; Yi et al., 2025; Xu et al., 2026; Li et al., 2025; Zhang et al., 2024b). This creates a *structural mismatch*: the edge has data but limited capability, while the cloud has capability but is constrained by privacy restrictions.

Recent work in privacy-preserving edge-cloud collaborative generation, detailed in Appendix A, has explored this mismatch (Hao et al., 2024; Zhan et al., 2026; Patel et al., 2024). *CoGenesis* uses cloud-generated sketches to guide an edge-side small language model (SLM) in personalized completion with private data (Zhang et al., 2024a). *LSRP* uses task-specific cloud guidelines within a leader-subordinate framework. (Zhang et al., 2025). More recently, *PRISM* profiles entity-level sensitivity at the edge, selects among cloud, edge, and collaborative execution modes, and applies adaptive local differential privacy before cloud-side semantic sketch generation (Zhan et al., 2026). While these methods show the promise of privacy-preserving edge-cloud collaboration, they remain largely *coarse-grained* and *single-pass*, failing to fully exploit the division-of-labor advantage of heterogeneous edge-cloud models (Pang et al., 2024; Zeng et al., 2025). This makes them less effective for complex queries requiring multi-step reasoning over mixed public and private knowledge (Khot et al., 2022; Trivedi et al., 2023; Besta et al., 2024).

Consider, for example, a complex financial query *"In the current environment of interest rates, should our company adjust the short-term debt structure?"*, whose resolution requires both private knowledge (*company debt*) and public knowledge (*interest rates*) across multiple reasoning steps,

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.

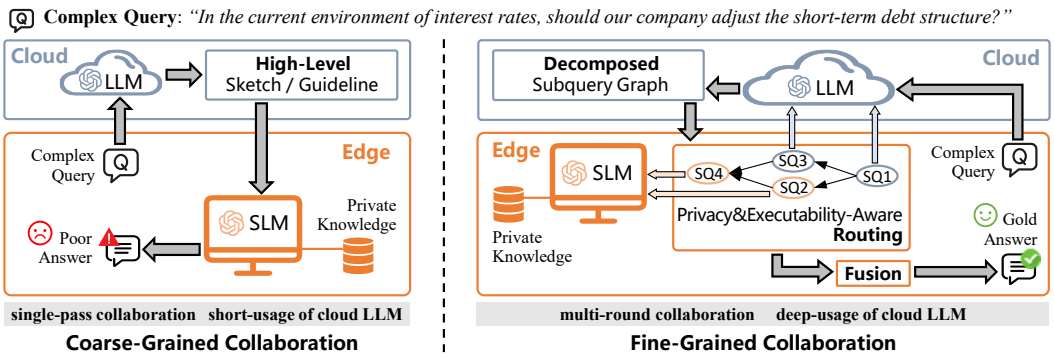


Figure 1: Motivation for fine-grained edge-cloud collaborative RAG. **Left**: existing methods treat the query as a whole and mainly use the cloud-side LLM as a sketch or guideline producer, with limited performance on complex queries due to the short usage of cloud-side LLM in subquery execution. **Right**: when a complex query is decomposed into a subquery graph and each subquery is appropriately assigned to the edge or cloud according to its privacy and capability requirements, higher answer quality can be achieved.

e.g., macroeconomic analysis → debt assessment → cost simulation → adjustment recommendation. As seen on the **left** of Fig. 1, even when the cloud provides promising sketches or guidelines, the edge-side SLM may still struggle to complete the entire query reliably due to the limited capability of edge-side SLMs. Consequently, a more effective solution is to more fully exploit the cloud LLM by offloading privacy-independent and edge-intractable subqueries for cloud-side execution. As shown on the **right** of Fig. 1, when a complex query is decomposed into dependent subqueries and the execution side of each subquery is assigned appropriately, the resulting answer quality is substantially better than that of coarse-grained edge-cloud collaboration. This observation points to *a promising direction*: fine-grained edge-cloud collaboration at the subquery level, but also faces a key challenge:

How to assign subquery to the edge or cloud appropriately, under privacy and capability constraints?

To address this challenge, we propose **SPEAR**, a Subquery-level Privacy- and Executability-Aware Routing method for edge-cloud collaborative RAG over complex queries. SPEAR first desensitizes the user query and lets the cloud decompose it into a subquery graph.The edge then estimates each subquery's privacy dependency and edge executability to jointly determine its execution side. Subqueries are executed in dependency order, and their outputs are fused at the edge into the final answer. Experiments on multiple privacy-sensitive domain datasets show that SPEAR consistently improves collaborative generation quality while preserving private knowledge protection. Our main contributions are as follows:

- We formulate privacy-constrained edge-cloud collaboration over complex multi-step RAG tasks, as a subquery-level execution-side assignment problem over a dependency-structured subquery graph.
- We propose a privacy dependency estimator that characterizes how strongly a subquery relies on private knowledge from both functional gain and structural necessity.
- We propose an edge executability estimator that evaluates whether a subquery can be reliably handled at the edge in terms of answer quality and reasoning consistency.
- We develop an edge-cloud routing and execution framework that jointly uses privacy dependency and edge executability for fine-grained collaborative generation, and validate its effectiveness on multiple domain datasets.

2 Problem Setup

2.1 Collaborative Generation Setting

We consider an edge-cloud collaborative generation setting under private knowledge constraints, as illustrated in Fig. 2. The edge has access to a private knowledge base $\mathcal{K}_{pri}$, hosts an SLM under

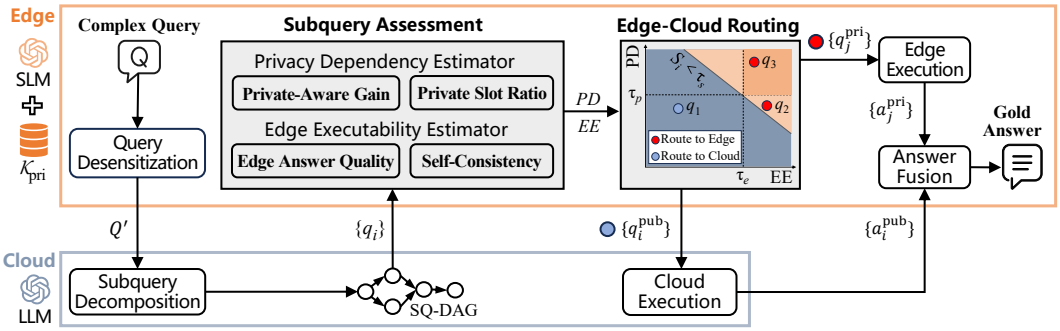


Figure 2: Edge-cloud collaborative generation under private knowledge constraints. The cloud decomposes the desensitized query, while the edge evaluates each subquery for privacy dependency and executability. Edge-cloud routing determines execution for each node, and the final answer is produced via edge-side fusion without exposing raw private knowledge.

edge-side resource constraints, and is responsible for privacy-sensitive local execution and final answer fusion. In contrast, the cloud has access to public knowledge and cloud-visible inputs, hosts an LLM, and is responsible for high-level planning and the execution of cloud-assigned subqueries.

Given a user query Q, the edge first applies desensitization and produces a transformed query Q', which is sent to the cloud without directly exposing raw private knowledge. The cloud then performs coarse-grained planning over Q' and decomposes it into a subquery graph (SQ-DAG). Formally, we denote the subquery graph by

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}), \tag{1}$$

where each node $v_i \in \mathcal{V}$ corresponds to a subquery q_i, and each directed edge $(v_j, v_i) \in \mathcal{E}$ indicates that resolving q_i depends on the output of q_j. Let $\mathrm{Pa}(i)$ denote the parent set of node v_i. For each node v_i, we define its parent context as

$$h_i = \{a_j \mid j \in \mathrm{Pa}(i)\}, \tag{2}$$

where a_j is the answer to subquery q_j. We further define an execution-side variable

$$z_i \in \{\texttt{edge}, \texttt{cloud}\}, \tag{3}$$

which indicates whether q_i is executed locally with access to $\mathcal{K}_{pri}$ or sent to the cloud for execution.

Our goal is to answer Q collaboratively by determining an execution-side variable z_i for each subquery q_i, while minimizing private knowledge exposure and maintaining high answer quality.

2.2 Privacy Threat Model

We assume that the edge is trusted and stores the private knowledge base $\mathcal{K}_{pri}$, while the cloud is *honest-but-curious*: it follows the protocol but may attempt to infer private knowledge from the information it receives, which is consistent with the commonly used semi-honest setting in secure machine learning systems (Bonawitz et al., 2017; Le et al., 2023).

Under this threat model, raw private knowledge must remain at the edge. The cloud observes the released query Q', the subquery graph, cloud requests and responses, and communication metadata, including message sizes, timing, ordering, and termination. The final answer remains local. Desensitization does not by itself imply statistical independence from private information. We therefore distinguish empirical leakage reduction from the formal guarantee of no additional disclosure beyond the initial public information, established for the certified execution variant.

3 Method

3.1 Method Overview

Building on the problem setup above, our method determines the execution side of each subquery through three tightly coupled components: **Privacy Dependency Estimator** (PDE), **Edge Executability Estimator** (EEE), and **Edge-Cloud Routing** (ECR), as illustrated in Fig. 2.

For each subquery q_i, we first estimate a privacy dependency score PD_i, which quantifies how strongly the subquery depends on edge-side private knowledge. We then estimate an edge executability score EE_i, which quantifies whether the edge-side SLM can reliably resolve the subquery. Finally, PD_i and EE_i are jointly used by Edge-Cloud Routing to determine the execution side.

The resulting node-level assignments enable privacy-aware collaborative execution over the SQ-DAG, followed by edge-side fusion for final answer generation.

3.2 Privacy Dependency Estimator

To identify subqueries that inherently depend on private knowledge, we estimate a privacy dependency score $\mathrm{PD}_i \in [0, 1]$ for each subquery q_i, using two complementary signals:

$$\mathrm{PD}_i = \alpha G_i^{\text{pri}} + (1 - \alpha) D_i^{\text{slot}}, \tag{4}$$

where G_i^{pri} is the *private-aware gain*, D_i^{slot} is the *private-slot ratio*, and $\alpha \in [0, 1]$.

Private-Aware Gain. We construct two counterfactual answers for q_i under the same parent context h_i: a private-aware answer a_i^{pri} obtained with access to $\mathcal{K}_{pri}$ and a private-agnostic answer $a_i^{\text{-pri}}$ generated without such access. We compare them using a lightweight utility function

$$U(a) = \lambda_1 \,\mathrm{Dir}(a) + \lambda_2 \,\mathrm{Spec}(a), \qquad \lambda_1 + \lambda_2 = 1, \tag{5}$$

where $\mathrm{Dir}(\cdot)$ measures whether the answer resolves the current subquery, and $\mathrm{Spec}(\cdot)$ measures whether it is specific enough for downstream continuation. The private-aware gain is defined as

$$G_i^{\text{pri}} = \max\big(0,\ U(a_i^{\text{pri}}) - U(a_i^{\text{-pri}})\big). \tag{6}$$

This term captures the functional gain brought by private access.

Private-Slot Ratio. We further estimate privacy dependency from the reasoning structure of q_i. Specifically, we elicit a minimal slot set

$$\mathcal{R}_i = \{r_{i1}, r_{i2}, \ldots, r_{iK}\}, \tag{7}$$

where each slot denotes a necessary information unit for resolving q_i, such as internal metrics, entity-specific states, decision constraints, or inherited intermediate facts. Let $\mathcal{R}_i^{\text{pri}} \subseteq \mathcal{R}_i$ denote the subset of slots that must be filled using private knowledge. We define

$$D_i^{\text{slot}} = \frac{\sum_{r \in \mathcal{R}_i} w(r)\, \mathbf{1}[r \in \mathcal{R}_i^{\text{pri}}]}{\sum_{r \in \mathcal{R}_i} w(r) + \varepsilon}, \tag{8}$$

where $w(r)$ is a lightweight importance score reflecting whether slot r is critical to the subquery answer, and ε is a small positive constant for numerical stability. This term captures how much of the reasoning structure is inherently private-dependent.

3.3 Edge Executability Estimator

To determine whether a subquery can be reliably resolved at the edge, we estimate an edge executability score $\mathrm{EE}_i \in [0, 1]$ for each subquery q_i. While PD_i captures *necessity*, EE_i captures *capability and stability*. We define

$$\mathrm{EE}_i = \gamma D_i^{\text{qual}} + (1 - \gamma) D_i^{\text{cons}}, \tag{9}$$

where $\gamma \in [0, 1]$, D_i^{qual} is the *edge answer quality*, and D_i^{cons} is the *edge self-consistency*.

Edge Answer Quality. Given an edge answer a_i^{edge} generated with access to $\mathcal{K}_{\text{pri}}$, we reuse the utility function in Eq. (5) and define

$$D_i^{\text{qual}} = U(a_i^{\text{edge}}). \tag{10}$$

This term captures whether the edge model can produce a direct and sufficiently specific answer for the current subquery.

Edge Self-Consistency. A subquery may appear answerable but still exhibit unstable reasoning. Inspired by self-consistency decoding (Wang et al., 2022), we sample the edge-side SLM M times with the same input and obtain $\{a_{i,1}^{\text{edge}}, a_{i,2}^{\text{edge}}, \ldots, a_{i,M}^{\text{edge}}\}$. We then compute a conclusion-level consistency score

$$D_i^{\text{cons}} = \text{Cons}\Big(\{a_{i,m}^{\text{edge}}\}_{m=1}^{M}\Big), \tag{11}$$

where $\text{Cons}(\cdot) \in [0,1]$ measures whether the sampled answers agree on the final conclusion of the subquery, e.g., the same decision, ranking, extracted value, or recommendation.

3.4 Edge-Cloud Routing

With PD_i and EE_i, we determine whether each subquery q_i should be executed at the edge or in the cloud, following two principles. First, it prioritizes *hard* privacy-first execution: if a subquery strongly depends on private knowledge and is executable at the edge, it should remain local. Second, for ambiguous cases, the joint score S_i provides a *soft* trade-off between privacy necessity and edge capability. Thus, we define a joint edge preference score

$$S_i = \eta\,\text{PD}_i + (1-\eta)\,\text{EE}_i, \tag{12}$$

where $\eta \in [0,1]$ balances privacy necessity against edge executability, and make the decision

$$z_i = \begin{cases} \texttt{edge}, & \text{if } \text{PD}_i \geq \tau_p \text{ and } \text{EE}_i \geq \tau_e \quad \textit{(hard rule)}, \\ \texttt{edge}, & \text{if } S_i \geq \tau_s \quad \textit{(soft score)}, \\ \texttt{cloud}, & \text{otherwise.} \end{cases} \tag{13}$$

where τ_p, τ_e, and τ_s are predefined thresholds.

3.5 Privacy Protection

Privacy-protected execution. Let A denote pre-existing public information, $X = (Q, \mathcal{K}_{\text{pri}})$ the protected input, and $\mathcal{R}$ the routing function in Eq. (13). For the locally computed, clipped scores $(s_i^{\text{p}}, s_i^{\text{e}}) = (\text{PD}_i, \text{EE}_i) \in [0,1]^2$, the edge uses

$$\widehat{s}_i^k = \text{clip}_{[0,1]}(s_i^k + \xi_i^k), \quad \xi_i^k \sim \text{Lap}(1/\epsilon_i^k), \quad k \in \{\text{p}, \text{e}\}, \tag{14}$$

and routes by $\widehat{z}_i = \mathcal{R}(\widehat{s}_i^{\text{p}}, \widehat{s}_i^{\text{e}})$. Here clipping projects onto $[0,1]$, and $\text{Lap}(b)$ is zero-mean Laplace noise of scale b, sampled independently for each invocation. Raw scores and unprotected decisions are not released.

For each outgoing message using unreleased private information, the edge samples $c \in \mathcal{C}_i(A,t)$ with probability proportional to $\exp(\epsilon_i^{\text{m}} v_i(X,t,c)/2)$, where t is the released transcript, $v_i \in [0,1]$ is a local relevance score, and the finite, nonempty candidate set depends only on (A,t). The initial query Q' uses the same mechanism with budget ϵ_0; it is not included in A. Messages derived solely from A and protected releases incur no additional privacy cost. Private parent answers remain local unless protected through this interface. All budgets and the maximum number of steps N are public. Observable timing, message lengths, retries, and termination depend only on A and protected releases, with no unprotected private-dependent overrides.

Privacy guarantee. Under these requirements, the full cloud transcript Y is ϵ_{tot}-differentially private across all protected inputs compatible with the same A, where $\epsilon_{\text{tot}} = \epsilon_0 + \sum_{i=1}^{N}(\epsilon_i^{\text{p}} + \epsilon_i^{\text{e}} + \epsilon_i^{\text{m}})$. For any finite protected attribute $S = f(X)$,

$$I(S; Y \mid A) \leq \epsilon_{\text{tot}} \tanh(\epsilon_{\text{tot}}/2). \tag{15}$$

Bounded score sensitivity, protected message sampling, and adaptive composition establish this guarantee without estimator-failure probabilities. Appendix B provides the complete assumptions, proof, and attribute-inference AUC bound.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Datasets. We evaluate SPEAR on privacy-augmented RAG datasets: *P-PubMedQA*, *P-FinQA*, and *P-MultiHopRAG*, derived from PubMedQA (Jin et al., 2019), FinQA (Chen et al., 2021), and MultiHopRAG (Tang & Yang, 2024) by adding private attributes (e.g., patient demographics and clinical history). Questions and reference answers are jointly rewritten to require private evidence. We retain only instances with answers supported by public and private evidence but incomplete or indeterminate without private evidence. Appendix C.1 details construction and human validation.

Baselines. We compare SPEAR 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.

Evaluation metrics, LLM/SLM configurations, and implementation details are provided in Appendices C.3, C.4, and C.5, respectively.

Table 1: Overall performance comparison between SPEAR, Cloud/Edge-Only and SOTA methods.

Method	P-PubMedQA					P-FinQA					P-MultiHopRAG				
	Accuracy			Privacy		Accuracy			Privacy		Accuracy			Privacy	
	F1↑	SM↑	LaJ↑	AUC↓	MI↓	F1↑	SM↑	LaJ↑	AUC↓	MI↓	F1↑	SM↑	LaJ↑	AUC↓	MI↓
LLM1+SLM1															
Cloud-Only	**.667**	**.762**	**.745**	1.0	1.0	**.741**	**.745**	**.795**	1.0	1.0	**.604**	**.787**	**.743**	1.0	1.0
RemoteRAG	.619	.661	.687	.687	.187	.684	.703	.748	.676	.221	.534	.691	.688	.683	.225
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	.445	.547	.531	**.500**	**.000**	.541	.615	.655	**.500**	**.000**	.458	.586	.547	**.500**	**.000**
SPEAR (*Ours*)	.657	.704	.717	.533	.116	.734	.739	.763	.538	.164	.587	.767	.724	.546	.175
LLM1+SLM2															
Cloud-Only	**.671**	**.692**	**.821**	1.0	1.0	**.659**	**.786**	**.794**	1.0	1.0	**.554**	**.751**	**.786**	1.0	1.0
RemoteRAG	.607	.649	.744	.679	.214	.585	.684	.722	.669	.194	.503	.684	.724	.649	.188
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	.488	.516	.574	**.500**	**.000**	.493	.524	.561	**.500**	**.000**	.442	.531	.563	**.500**	**.000**
SPEAR (*Ours*)	.632	.667	.798	.567	.124	.626	.754	.778	.566	.113	.514	.723	.761	.560	.125
LLM2+SLM1															
Cloud-Only	**.698**	**.733**	**.801**	1.0	1.0	**.760**	**.792**	**.797**	1.0	1.0	**.584**	**.809**	**.815**	1.0	1.0
RemoteRAG	.612	.649	.714	.662	.195	.675	.692	.713	.647	.228	.512	.703	.757	.652	.217
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	.465	.534	.594	**.500**	**.000**	.491	.552	.597	**.500**	**.000**	.442	.589	.637	**.500**	**.000**
SPEAR (*Ours*)	.661	.680	.762	.534	.112	.721	.748	.766	.547	.155	.536	.764	.782	.537	.151
LLM2+SLM2															
Cloud-Only	**.647**	**.687**	**.825**	1.0	1.0	**.711**	**.782**	**.806**	1.0	1.0	**.543**	**.752**	**.793**	1.0	1.0
RemoteRAG	.593	.642	.773	.623	.178	.642	.717	.736	.652	.197	.501	.692	.716	.643	.218
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	.513	.582	.638	**.500**	**.000**	.532	.578	.616	**.500**	**.000**	.433	.559	.619	**.500**	**.000**
SPEAR (*Ours*)	.618	.662	.802	.569	.118	.684	.752	.781	.537	.113	.518	.716	.757	.545	.178

4.2 OVERALL PERFORMANCE COMPARISON

Table 1 shows that SPEAR achieves the best balance between answer quality and privacy protection across all datasets and model settings. SPEAR outperforms all privacy-preserving baselines in F1,

SM, and LaJ, while remaining close to the Cloud-Only upper bound, indicating that fine-grained edge-cloud collaboration preserves most quality advantages of full cloud execution. Meanwhile, SPEAR ranks second in privacy metrics, only behind Edge-Only, and yields lower AUC and MI than PRISM, LSRP, CoGenesis, MINIONS, and RemoteRAG. Although RemoteRAG achieves second-best answer quality, it incurs higher privacy leakage. These verify SPEAR's superior privacy-utility trade-off by offloading only privacy-insensitive and edge-intractable subqueries to the cloud.

Table 1 further suggests that model compatibility matters. Under the same cloud LLM1, same-family LLM1+SLM1 generally achieves better F1 and lower AUC than cross-family LLM1+SLM2, indicating better local completion and privacy control. When fixing SLMs, the closed-source LLM2 improves LaJ over open-source LLM1, while gains on F1, SM, and privacy are less consistent. More LLM+SLM settings in Appendix D further confirm SPEAR's generality.

4.3 Privacy Leakage Analysis across Transmission Stages

Table 2: Privacy leakage analysis of SPEAR across four transmission stages.

Metric Total	Model	P-PubMedQA				P-FinQA				P-MultiHopRAG			
		q2c	sq2e	pq2c	pa2e	q2c	sq2e	pq2c	pa2e	q2c	sq2e	pq2c	pa2e
AUC↓	LLM1+SLM1	.548	.511	.542	.531	.557	.525	.541	.529	.551	.539	.549	.545
	LLM1+SLM2	.583	.554	.569	.561	.592	.556	.594	.523	.574	.551	.568	.547
	LLM2+SLM1	.552	.517	.551	.518	.563	.552	.556	.517	.557	.527	.541	.524
	LLM2+SLM2	.561	.548	.599	.567	.552	.542	.534	.521	.557	.537	.554	.539
MI↓	LLM1+SLM1	.033	.023	.032	.028	.049	.031	.045	.039	.051	.038	.047	.039
	LLM1+SLM2	.038	.026	.031	.029	.058	.042	.047	.038	.044	.035	.042	.038
	LLM2+SLM1	.035	.024	.031	.022	.047	.035	.042	.031	.043	.034	.039	.035
	LLM2+SLM2	.038	.031	.030	.019	.055	.046	.048	.041	.048	.042	.044	.041

Table 2 shows that SPEAR maintains low privacy leakage across transmission stages, where q2c, sq2e, pq2c, and pa2e denote Q', q_i, q_i^{pub}, and a_i^{pub} across corresponding edge-cloud directions. Leakage is generally higher for the uploaded desensitized query (*q2c*) and public subquery (*pq2c*) than for the cloud-returned subquery (*sq2e*) and public answer (*pa2e*), suggesting that residual risk mainly lies in edge-to-cloud transmission. Moreover, lower leakage is generally observed under same-family pairings and stronger closed-source cloud models. Additional results in Appendix E confirm SPEAR's robustness against stronger attackers and show that major leakage sources stem from implicit contextual cues rather than explicit private terms.

4.4 Ablation Study on PDE, EEE, and ECR

Table 3 verifies that PDE, EEE, and ECR are all necessary for SPEAR. The full model (*Line 1*) achieves the best accuracy and privacy, while removing any component causes clear degradation. PDE mainly improves privacy control, as its removal (*Lines 2,6,7*) leads to the largest privacy increase. EEE enhances the reliability of local execution and benefits both utility and privacy. ECR has the largest impact on answer quality, highlighting the importance of routing. When both PDE and EEE are removed, retaining only ECR causes severe degradation, showing that routing must rely on accurate privacy and executability signals.

Table 3: Ablation on key modules.

PDE	EEE	ECR	Accuracy			Privacy	
			F1	SM	LaJ	AUC	MI
✓	✓	✓	.657	.704	.717	.533	.116
✗	✓	✓	.584	.609	.664	.668	.209
✓	✗	✓	.573	.597	.675	.585	.162
✓	✓	✗	.534	.574	.627	.593	.176
✓	✗	✗	.498	.545	.562	.648	.174
✗	✓	✗	.491	.538	.554	.679	.218
✗	✗	✓	.265	.302	.341	.898	.459

Table 4 shows that both private-aware gain G_i^{pri} and private-slot ratio D_i^{slot} are important in PDE. The full version achieves the best accuracy and privacy, while removing either causes degradation. Among them, D_i^{slot} has a larger privacy effect, highlighting structural private-slot estimation, whereas G_i^{pri} provides complementary functional evidence for identifying privacy-dependent subqueries.

Table 4: Ablation on G_i^{pri} and D_i^{slot} indicators in PDE.

G_i^{pri}	D_i^{slot}	Accuracy			Privacy	
		F1	SM	LaJ	AUC	MI
✓	✓	.657	.704	.717	.533	.116
✓	✗	.611	.635	.678	.589	.154
✗	✓	.624	.644	.701	.567	.137
✗	✗	.584	.609	.664	.668	.209

Table 5: Ablation on D_i^{qual} and D_i^{cons} indicators in EEE.

D_i^{qual}	D_i^{cons}	Accuracy			Privacy	
		F1	SM	LaJ	AUC	MI
✓	✓	.657	.704	.717	.533	.116
✓	✗	.625	.665	.691	.558	.133
✗	✓	.587	.610	.685	.574	.156
✗	✗	.573	.597	.675	.585	.162

Table 6: Ablation on hard and soft routing strategies in ECR.

z_i Strategy		Accuracy			Privacy	
Hard	Soft	F1	SM	LaJ	AUC	MI
✓	✓	.657	.704	.717	.533	.116
✓	✗	.547	.581	.658	.568	.139
✗	✓	.568	.607	.675	.549	.127
✗	✗	.534	.574	.627	.593	.176

Table 5 verifies that both the answer quality D_i^{qual} and self-consistency D_i^{cons} contribute to EEE. The full model achieves the best answer quality and privacy protection, while removing either one causes degradation. Among them, D_i^{qual} has a larger impact, indicating that answer quality is the primary signal for edge executability, whereas D_i^{cons} provides complementary evidence on execution stability.

Table 6 shows that both the hard rule and the soft score are important in the routing strategy of Eq. (13). The full version achieves the best answer quality and privacy protection, while removing either one causes clear degradation. In particular, removing the soft score leads to a larger drop, indicating its importance in resolving ambiguous routing cases, whereas the hard rule provides necessary privacy-first constraints for strongly privacy-dependent subqueries.

More ablation results in Appendix F further confirm the roles of SPEAR's modules and indicators.

Table 7: Edge-cloud routing accuracy.

Datasets	Edge Acc.	Cloud Acc.	Acc.	AUC	MI
P-PubMedQA	.974	.940	.957	.533	.116
P-FinQA	.962	.922	.942	.538	.164
P-MultiHopRAG	.953	.923	.938	.546	.175

Table 8: Hard rule sensitivity in ECR.

Method	Accuracy			Privacy	
	F1	SM	LaJ	AUC	MI
and	.657	.704	.717	.533	.116
or	.634	.682	.689	.557	.142

Table 7 reports the routing accuracy of ECR against human-annotated ground-truth labels. ECR achieves high accuracy on all three datasets, all above 0.9, showing that it can reliably distinguish edge-routed from cloud-routed subqueries. In addition, edge-side routing is more accurate than cloud-side routing, empirically supporting the privacy-first hard rule in ECR.

4.5 Overhead Analysis

Table 9: Token costs and latency across baselines under LLM1+SLM1.

Method	P-PubMedQA			P-FinQA			P-MultiHopRAG		
	Lat.(s)	I. Tokens	O. Tokens	Lat.(s)	I. Tokens	O. Tokens	Lat.(s)	I. Tokens	O. Tokens
Cloud-Only	52.67	231	4934	86.54	136	8587	78.51	265	7024
RemoteRAG	58.71	328	5354	93.22	264	8799	86.56	289	7130
MINIONS	69.32	251	6194	110.39	156	9965	118.65	274	9259
CoGenesis	64.36	276	5854	99.64	172	9539	106.68	287	8827
LSRP	62.25	249	5459	98.47	148	9117	101.29	269	8301
PRISM	59.56	244	5968	100.69	179	9728	109.37	283	8948
Edge-Only	7.29	231	4235	9.64	136	6924	8.31	265	6037
SPEAR (*Ours*)	65.13	307	6566	104.68	239	10443	108.57	282	10883

Table 9 shows that SPEAR incurs higher latency and token cost than other edge-cloud baselines, under the same LLM1+SLM1 setting as in Table 1, which is expected given its fine-grained multi-stage collaboration. Nevertheless, its overhead remains moderate relative to PRISM, LSRP, CoGenesis, and MINIONS, while delivering clearly better privacy-utility trade-offs. Table 10 further shows that the dominant overhead comes from cloud-side decomposition and edge-side subquery execution, while desensitization, PDE, EEE, ECR, and fusion incur only minor latency. In addition, total latency increases smoothly as edge capability decreases from Nvidia 3080 to 2080Ti and 2080, suggesting that SPEAR introduces structured and predictable overhead rather than excessive system complexity. Similar observations can also be seen under mobile edge settings, as reported in Appendix G.

Table 10: Stage-wise latency of SPEAR under different edge capacities, measured in seconds.

Edge	Stage										Total
	Desen.	Q2Cloud	Decom.	SQ2Edge	PDE	EEE	ECR	SQExecution	Fusion	Answer	
3080	0.31	0.78	45.81	0.73	0.32	0.27	0.35	14.54	1.45	0.57	65.13
2080Ti	0.49	1.24	46.28	1.05	0.47	0.37	0.51	15.31	1.98	0.87	68.57
2080	0.65	1.69	46.97	1.59	0.73	0.58	0.74	15.39	2.88	1.06	72.28

4.6 Sensitivity Analysis

Sensitivity of function implementations. Table 11 analyzes the sensitivity of SPEAR to different function implementations. For subquery answer directness, i.e., $\mathrm{Dir}(\cdot)$, the keyword-semantic score (KSS) clearly outperforms KL divergence (KLD) (Zhang et al., 2023). For answer specificity in downstream continuation, i.e., $\mathrm{Spec}(\cdot)$, sliding window similarity (SWS) achieves the best performance. For slot criticality estimation, i.e., $w(r)$, TF-IDF contribution (TIC) (Wu et al., 2008) performs best. For conclusion-level agreement among sampled answers, i.e., $\mathrm{Cons}(\cdot)$, N-gram coverage (NG) (Diao et al., 2020) yields the best results.

Sensitivity of hard rule implementations. Table 8 compares two implementations of the hard-rule strategy in Eq. (13) and shows that the "*and*" rule outperforms the "*or*" rule in ECR. It yields both higher answer quality and lower privacy leakage, suggesting that stricter privacy-first routing is more effective for selecting truly edge-suitable subqueries while avoiding unnecessary privacy exposure.

Table 11: The sensitivity of SPEAR to function implementations in P-PubMedQA.

Function	Meaning	Method	Description	Accuracy			Privacy	
				F1	SM	LaJ	AUC	MI
$\mathrm{Dir}(\cdot)$	answer directness to the subquery	KSS	Keyword-Semantic Score	.657	.704	.717	.533	.116
		KLD	KL Divergence	.638	.684	.703	.549	.138
$\mathrm{Spec}(\cdot)$	answer specificity for downstream reasoning	SWS	Sliding Window Similarity	.657	.704	.717	.533	.116
		CSD	Concrete Sentence Density	.642	.688	.701	.542	.124
$w(r)$	slot importance to subquery answer	TIC	TF-IDF Contribution	.657	.704	.717	.533	.116
		KC	Keyword Coverage	.643	.681	.707	.548	.139
$\mathrm{Cons}(\cdot)$	conclusion agreement among sampled answers	NG	N-Gram Coverage	.657	.704	.717	.533	.116
		SS	Semantic Similarity	.644	.689	.703	.556	.142
		KC	Keyword Coverage	.635	.672	.691	.568	.161

Additional sensitivity results in Appendices H and I further confirm the robustness of SPEAR across different implementations and hyperparameter settings.

We also provide a **case study** in Appendix J to detail the execution of SPEAR's key modules, and a **real-world mobile edge device** demo in Appendix K to demonstrate its practical applicability.

5 Conclusion

We proposed SPEAR, a privacy- and executability-aware subquery routing framework for edge-cloud collaborative RAG. SPEAR 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 RAG datasets show that SPEAR 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 L.

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 support reproducibility, the anonymized repository linked in the abstract provides the implementation of SPEAR, experiment configurations, prompt templates, random seeds, and evaluation scripts.

the repository provides scripts and instructions for obtaining the corresponding public datasets from their original providers.

For each experiment, we report the cloud-side and edge-side model identifiers, software and hardware environments, decoding settings, routing hyperparameters, and the number of sampled edge-side generations. For certified execution, we additionally release the enforced routing cap N, the privacy-budget schedule $(\epsilon_0, \epsilon_i^p, \epsilon_i^e, \epsilon_i^m)$, candidate-message construction procedure, and metadata-control policy. The repository also includes scripts for answer-quality evaluation, finite-sample attacker-based AUC evaluation, and the explicitly specified empirical mutual-information estimator. The latter is reported separately from the theoretical target quantity $I(S;Y)$.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Lameck Mbangula Amugongo, Pietro Mascheroni, Steven Brooks, Stefan Doering, and Jan Seidel. Retrieval augmented generation for large language models in healthcare: A systematic review. *PLOS Digital Health*, 4(6):e0000877, 2025.

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.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

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.

Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1175–1191, 2017.

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.

Shizhe Diao, Jiaxin Bai, Yan Song, Tong Zhang, and Yonggang Wang. Zen: Pre-training chinese text encoder enhanced by n-gram representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4729–4740, 2020.

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.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 6491–6501, 2024.

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.

Team Glm, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

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.

Junqing Le, Di Zhang, Xinyu Lei, Long Jiao, Kai Zeng, and Xiaofeng Liao. Privacy-preserving federated learning with malicious clients and honest-but-curious servers. *IEEE Transactions on Information Forensics and Security*, 18:4329–4344, 2023.

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.

Fatemehsadat Mireshghallah, Kartik Goyal, Archit Uniyal, Taylor Berg-Kirkpatrick, and Reza Shokri. Quantifying privacy risks of masked language models using membership inference attacks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 8332–8347, 2022.

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.

Pratyush Patel, Esha Choukse, Chaojie Zhang, Aashaka Shah, Íñigo Goiri, Saeed Maleki, and Ricardo Bianchini. Splitwise: Efficient generative llm inference using phase splitting. In *2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA)*, pp. 118–132. 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.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

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.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Zhiqi Wang, Chengyu Zhang, Yuetian Chen, Nathalie Baracaldo, Swanand R Kadhe, and Lei Yu. Membership inference attacks as privacy tools: Reliability, disparity and ensemble. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1724–1738, 2025.

Ho Chung Wu, Robert Wing Pong Luk, Kam Fai Wong, and Kui Lam Kwok. Interpreting tf-idf term weights as making relevance decisions. *ACM Transactions on Information Systems (TOIS)*, 26(3): 1–37, 2008.

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.

Anbang Xu, Tan Yu, Min Du, Pritam Gundecha, Yufan Guo, Xinliang Zhu, May Wang, Ping Li, and Xinyun Chen. Generative ai and retrieval-augmented generation (rag) systems for enterprise. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 5599–5602, 2024.

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.

Minrui Xu, Dusit Niyato, Hongliang Zhang, Jiawen Kang, Zehui Xiong, Shiwen Mao, and Zhu Han. Cached model-as-a-resource: Provisioning large language model agents for edge intelligence in space-air-ground integrated networks. *IEEE Transactions on Networking*, 34:2850 – 2864, 2026.

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.

Xiaohan Yu, Pu Jian, and Chong Chen. Tablerag: A retrieval augmented generation framework for heterogeneous document reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)*, 2025. URL `https://aclanthology.org/2025.emnlp-main.710/`.

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.

Yufeng Zhang, Jialu Pan, Li Ken Li, Wanwei Liu, Zhenbang Chen, Xinwang Liu, and Ji Wang. On the properties of kullback-leibler divergence between multivariate gaussian distributions. *Advances in neural information processing systems*, 36:58152–58165, 2023.

Suifeng Zhao, Zhuoran Jin, Sujian Li, and Jun Gao. Finragbench-v: A benchmark for multimodal rag with visual citation in the financial domain. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 4215–4249, 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.

Rongzhi Zhu, Xiangyu Liu, Zequn Sun, Yiwei Wang, and Wei Hu. Mitigating lost-in-retrieval problems in retrieval augmented multi-hop question answering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025, Long Papers)*, 2025. URL https://aclanthology.org/2025.acl-long.1089/.

A Related Work

Edge-Cloud Collaborative Generation. Recent studies have explored edge-cloud collaboration to combine the reasoning capability of cloud-side LLMs with the privacy and locality advantages of edge-side SLMs. One line of work is motivated by privacy-preserving generation. CoGenesis integrates cloud-hosted large models with local small models for secure context-aware instruction following, and introduces sketch-based and logits-based collaboration mechanisms to keep sensitive context local (Zhang et al., 2024a). LSRP further studies privacy-preserving cloud-device collaboration through a leader-subordinate retrieval framework, where the cloud provides task-specific guidance and the edge-side SLM uses local private data for personalized generation (Zhang et al., 2025). PRISM introduces a privacy-aware routing framework that profiles entity-level sensitivity at the edge, selects among cloud, edge, and collaborative execution modes, and applies adaptive local differential privacy before cloud-side semantic sketch generation (Zhan et al., 2026). *These methods demonstrate the feasibility of privacy-aware edge-cloud generation, but they mainly follow coarse-grained guidance or single-pass collaboration.*

Another line of work focuses on reducing cloud cost, communication cost, or inference latency. MINIONS studies cost-efficient collaboration between on-device and cloud language models, where the remote model decomposes tasks into easier subtasks over shorter context chunks and the local model executes them in parallel (Narayan et al., 2025). Hybrid SLM-LLM proposes token-level edge-cloud collaborative inference, where an edge-side SLM selectively collaborates with a cloud-side LLM to reduce latency and improve efficiency (Hao et al., 2024). CE-LSLM introduces edge-cloud collaboration between LLMs and SLMs with semantic-level intermediate-state sharing, KV-cache reuse, and parallel scheduling to improve latency and communication efficiency (Patel et al., 2024). EdgeShard further explores collaborative LLM inference by partitioning model shards across heterogeneous edge and cloud resources to reduce latency and improve throughput (Zhang et al., 2024b). *Different from these efficiency-oriented methods, SPEAR focuses on privacy-constrained RAG and performs subquery-level routing based on both privacy dependency and edge executability.*

Privacy-Preserving RAG and Private Inference. Privacy-preserving RAG aims to leverage external knowledge while preventing sensitive queries, private documents, or retrieved contexts from being exposed to untrusted cloud services. RemoteRAG protects user-query privacy in cloud RAG through perturbed embeddings and restricted retrieval with a distance-based differential privacy formulation (Cheng et al., 2025). Private inference studies commonly consider semi-honest cloud settings, where the server may infer private information from received inputs or intermediate representations (Pang et al., 2024). Other work studies output-side leakage from retrieved knowledge and applies differential privacy to reduce sensitive data exposure (Zeng et al., 2025). *In contrast, SPEAR keeps the private knowledge base on the edge and prevents the cloud from directly observing the raw query, while further controlling the leakage risks* introduced by intermediate communication artifacts during subquery decomposition, planning, and edge-cloud execution.

Query Decomposition and Multi-Step Reasoning. Complex knowledge-intensive tasks often require decomposing a user query into multiple intermediate reasoning steps. Decomposed Prompting solves complex tasks by breaking them into simpler sub-tasks that can be delegated to specialized prompts, models, or external modules (Khot et al., 2022). IRCoT interleaves retrieval with chain-of-thought reasoning, showing that retrieval and reasoning can benefit from being performed iteratively in multi-step question answering (Trivedi et al., 2023). Graph of Thoughts further generalizes reasoning into graph-structured intermediate states, enabling more flexible reasoning paths than linear chains (Besta et al., 2024). These works motivate the use of structured subqueries for complex RAG tasks. *SPEAR differs from them by studying not how to decompose a query, but where each decomposed subquery should be executed under privacy and edge-capability constraints.*

B Privacy Guarantees and Proofs

This appendix establishes full-transcript privacy guarantees for SPEAR through protected score and message interfaces. We specify the execution requirements, derive mutual-information and attribute-inference bounds, and provide their proofs.

B.1 Protected Inputs and Observable Transcript

Let $X = (Q, \mathcal{K}_{\text{pri}})$ contain the original query and private knowledge. Let A denote information available to the cloud before the protocol, including public knowledge, public protocol configuration, and auxiliary background information. The newly released query Q' is not included in A; it is a protocol output whose privacy cost must be accounted for.

Let T be the complete cloud-visible transcript, including the released query, SQ-DAG, requests and responses, observable routing behavior, and communication metadata. We write $Y = (A, T)$ for the full cloud view. Private edge memory, local intermediate answers, and the final answer are not directly observable by the cloud.

We compare arbitrary admissible private inputs x, x' compatible with the same public information $A = a$. Thus, the guarantee protects the entire specified private input, rather than only a single-record modification. A protected attribute is a finite-valued variable $S = f(X)$, such as a private clinical or financial attribute. No independence between A and S is assumed unless explicitly stated.

All logarithms are natural, so mutual information and entropy are measured in nats.

B.2 Protected Execution Requirements

The guarantees below assume that the following score, message, and metadata restrictions are enforced.

Protected scores. Let $\mathcal{R}(p, e)$ denote the hard/soft routing function in Eq. (13). The edge computes the original PDE and EEE scores and clips them to $[0, 1]$. Before these scores influence cloud-visible execution, the edge forms

$$\begin{aligned} \widehat{\mathrm{PD}}_i &= \mathrm{clip}_{[0,1]}\left(\mathrm{PD}_i + \xi_i^{\mathrm{p}}\right), \quad \xi_i^{\mathrm{p}} \sim \mathrm{Lap}\left(1/\epsilon_i^{\mathrm{p}}\right), \\ \widehat{\mathrm{EE}}_i &= \mathrm{clip}_{[0,1]}\left(\mathrm{EE}_i + \xi_i^{\mathrm{e}}\right), \quad \xi_i^{\mathrm{e}} \sim \mathrm{Lap}\left(1/\epsilon_i^{\mathrm{e}}\right), \end{aligned} \tag{16}$$

where $\mathrm{clip}_{[0,1]}(v) = \min\{1, \max\{0, v\}\}$ and $\mathrm{Lap}(b)$ denotes zero-mean Laplace noise with scale b. The budgets $\epsilon_i^{\mathrm{p}}, \epsilon_i^{\mathrm{e}} > 0$ are publicly specified. Each invocation uses fresh noise independent of private inputs and previous randomness.

The operational decision is

$$\widehat{z}_i = \mathcal{R}\left(\widehat{\mathrm{PD}}_i, \widehat{\mathrm{EE}}_i\right). \tag{17}$$

Neither raw scores nor unprotected routing decisions are released. Any additional private-dependent decision affecting cloud-visible execution must also pass through a protected interface.

Protected message construction. A cloud request based only on A and previously protected releases is post-processing and incurs no additional privacy cost. A request using unreleased private information, including private parent answers, requires a protected message mechanism.

For a fixed public transcript prefix t, let $\mathcal{C}_i(A, t)$ be a finite nonempty candidate-message set constructed exclusively from A and t. This set must be identical for all private inputs compatible with the same (A, t). For example, candidates may be generated from public task templates or public knowledge. Private-dependent candidate generation or filtering is not permitted outside a separately protected mechanism.

Let $v_i(X, t, c) \in [0, 1]$ be a local relevance score for candidate c. The message mechanism samples

$$\Pr(M_i = c \mid X, A, t) = \frac{\exp\left(\epsilon_i^{\mathrm{m}} v_i(X, t, c)/2\right)}{\sum_{c' \in \mathcal{C}_i(A,t)} \exp\left(\epsilon_i^{\mathrm{m}} v_i(X, t, c')/2\right)}. \tag{18}$$

The score may use private information, but its range is bounded regardless of its estimation accuracy. No raw private-dependent prompt is transmitted alongside the sampled message.

The initial query Q' is released using the same construction with budget ϵ_0 and a candidate set determined only by A. Alternatively, a release that is genuinely a function only of A has zero privacy cost. Heuristic desensitization alone does not establish this zero-cost condition.

Dependency-aware continuation. SPEAR continues to process subqueries in dependency order. Private parent answers may be used for local estimation, execution, and fusion. If they influence an outbound request, that request must satisfy the protected-message requirement above. A local fallback or abstention is permitted, but any additional cloud call requires privacy accounting. Cloud responses are generated from the cloud's existing information and protected requests; no separate channel provides access to the private state.

Metadata and budget control. The protocol has a publicly specified upper bound N on routing steps and a public budget schedule. Requests, retries, and termination must not expose additional private-dependent decisions outside the protected interfaces. Message timing and size must depend only on public information and protected releases, or be masked by an appropriate schedule and padding policy. In particular, private computation time must not directly determine an observable request time or completion signal.

The absence of a cloud request may reveal $\widehat{z}_i$, but must not reveal an additional decision based on raw scores. Unused steps consume no additional information, and execution cannot exceed the public budget by issuing unaccounted private-dependent requests.

B.3 Transcript Privacy and Metric Bounds

Define the total privacy-budget upper bound

$$\epsilon_{\mathrm{tot}} = \epsilon_0 + \sum_{i=1}^{N} \left(\epsilon_i^{\mathrm{p}} + \epsilon_i^{\mathrm{e}} + \epsilon_i^{\mathrm{m}}\right), \tag{19}$$

where $\epsilon_i^{\mathrm{m}} = 0$ for a message determined solely by public information and preceding protected releases.

Proposition 1: Full-transcript privacy. Under the requirements in Appendix B.2, for every measurable transcript event $\mathcal{B}$ and every pair of compatible private inputs x, x',

$$\Pr(Y \in \mathcal{B} \mid X = x, A = a) \le e^{\epsilon_{\mathrm{tot}}} \Pr(Y \in \mathcal{B} \mid X = x', A = a). \tag{20}$$

This guarantee permits adaptive scoring and execution along the SQ-DAG.

Corollary 1: Mutual information. Let

$$c(\epsilon) = \epsilon \tanh(\epsilon/2). \tag{21}$$

For every finite-valued protected attribute $S = f(X)$,

$$I(S; Y \mid A) \le \min\left\{H(S \mid A), c(\epsilon_{\mathrm{tot}})\right\}, \tag{22}$$

and

$$I(S; Y) \le I(S; A) + \min\left\{H(S \mid A), c(\epsilon_{\mathrm{tot}})\right\}. \tag{23}$$

These bounds concern mutual information itself, not an attacker-dependent estimate of it.

Corollary 2: Attribute-inference AUC. Suppose $S \in \{0, 1\}$. For fixed $A = a$ with both classes having positive conditional probability, let $\mathrm{AUC}_a^\star(Y)$ denote the optimal population AUC among attribute-inference scoring rules under this conditional distribution. Using standard half-credit for tied scores,

$$\mathrm{AUC}_a^\star(Y) \le \frac{e^{\epsilon_{\mathrm{tot}}}}{1 + e^{\epsilon_{\mathrm{tot}}}}. \tag{24}$$

If $A \perp S$, the same upper bound applies to the unconditional optimal AUC. Without this additional condition, pre-existing attribute information in A cannot be ignored.

B.4 Concrete Privacy-Budget Configuration

For certified SPEAR execution, we enforce a public cap of $N = 7$ routing steps per query. This is a protocol-level cap rather than an observed average: after the third routing step, all remaining

nodes are executed locally at the edge. We use the following fixed privacy-budget schedule for every protected query:

$$\epsilon_0 = 0.10, \qquad \epsilon_i^p = 0.10, \qquad \epsilon_i^e = 0.10, \qquad \epsilon_i^m = 0.10, \quad i \in \{1, 2, 3\}. \tag{25}$$

Here, ϵ_0 protects the initial release Q', ϵ_i^p and ϵ_i^e protect the privacy-dependency and edge-executability scores at routing step i, respectively, and ϵ_i^m protects a private-dependent outgoing message. If an outgoing message is determined solely by A and preceding protected releases, we set $\epsilon_i^m = 0$ for that step. Thus, the following quantity is a worst-case per-query budget:

$$\epsilon_{\text{tot}} = \epsilon_0 + \sum_{i=1}^{N} (\epsilon_i^p + \epsilon_i^e + \epsilon_i^m) = 0.10 + 3(0.10 + 0.10 + 0.10) = 1.00. \tag{26}$$

Under this configuration, Proposition 1 gives 1.00-differential privacy for the complete cloud-visible transcript. Consequently, for every finite protected attribute $S = f(X)$,

$$I(S; Y \mid A) \le \min\{H(S \mid A),\ 1.00 \tanh(1.00/2)\} = \min\{H(S \mid A),\ 0.4621\} \text{ nats.} \tag{27}$$

The corresponding unconditional quantity satisfies

$$I(S; Y) \le I(S; A) + 0.4621 \quad \text{nats.} \tag{28}$$

For every fixed public context $A = a$, the optimal population attribute-inference AUC is bounded by

$$\text{AUC}_a^\star(Y) \le \frac{\exp(1.00)}{1 + \exp(1.00)} = 0.7311. \tag{29}$$

These are certified worst-case bounds on the complete transcript, not estimates of finite-sample attacker AUC or empirical mutual information. Any reported empirical MI estimator and attacker-based AUC should therefore be labeled as diagnostic measurements and reported separately from the bounds above. For a different enforced routing cap $N_{\max}$, the same schedule yields

$$\epsilon_{\text{tot}}(N_{\max}) = 0.10 + 0.30 N_{\max}. \tag{30}$$

B.5 Concrete Privacy-Budget Configuration

For certified SPEAR execution, we enforce a public cap of $N = 3$ routing steps per query. This is a protocol-level cap rather than an observed average: after the third routing step, all remaining nodes are executed locally at the edge. We use the following fixed privacy-budget schedule for every protected query:

$$\epsilon_0 = 0.10, \qquad \epsilon_i^p = 0.10, \qquad \epsilon_i^e = 0.10, \qquad \epsilon_i^m = 0.10, \quad i \in \{1, 2, 3\}. \tag{31}$$

Here, ϵ_0 protects the initial release Q', ϵ_i^p and ϵ_i^e protect the privacy-dependency and edge-executability scores at routing step i, respectively, and ϵ_i^m protects a private-dependent outgoing message. If an outgoing message is determined solely by A and preceding protected releases, we set $\epsilon_i^m = 0$ for that step. Thus, the following quantity is a worst-case per-query budget:

$$\epsilon_{\text{tot}} = \epsilon_0 + \sum_{i=1}^{N} (\epsilon_i^p + \epsilon_i^e + \epsilon_i^m) = 0.10 + 3(0.10 + 0.10 + 0.10) = 1.00. \tag{32}$$

Under this configuration, Proposition 1 gives 1.00-differential privacy for the complete cloud-visible transcript. Consequently, for every finite protected attribute $S = f(X)$,

$$I(S; Y \mid A) \le \min\{H(S \mid A),\ 1.00 \tanh(1.00/2)\} = \min\{H(S \mid A),\ 0.4621\} \text{ nats.} \tag{33}$$

The corresponding unconditional quantity satisfies

$$I(S; Y) \le I(S; A) + 0.4621 \quad \text{nats.} \tag{34}$$

For every fixed public context $A = a$, the optimal population attribute-inference AUC is bounded by

$$\text{AUC}_a^\star(Y) \le \frac{\exp(1.00)}{1 + \exp(1.00)} = 0.7311. \tag{35}$$

These are certified worst-case bounds on the complete transcript, not estimates of finite-sample attacker AUC or empirical mutual information. Any reported empirical MI estimator and attacker-based AUC should therefore be labeled as diagnostic measurements and reported separately from the bounds above. For a different enforced routing cap $N_{\max}$, the same schedule yields

$$\epsilon_{\text{tot}}(N_{\max}) = 0.10 + 0.30 N_{\max}. \tag{36}$$

B.6 PROOFS

Step 1: Privacy of score releases. Fix a public transcript prefix. For any clipped score $s(X) \in [0, 1]$ and compatible inputs x, x',

$$|s(x) - s(x')| \leq 1. \tag{37}$$

Let $W = s(X) + \mathrm{Lap}(1/\epsilon)$. For any output value w,

$$\begin{aligned} \frac{p(w \mid x)}{p(w \mid x')} &= \exp\left(\epsilon\left[|w - s(x')| - |w - s(x)|\right]\right) \\ &\leq \exp\left(\epsilon|s(x) - s(x')|\right) \\ &\leq e^{\epsilon}. \end{aligned} \tag{38}$$

Clipping the noisy score is post-processing and preserves this inequality. Therefore, the protected PDE and EEE scores have privacy budgets ϵ_i^{p} and ϵ_i^{e}, respectively.

The argument also covers randomized estimators and private-dependent local states. The bounded-range inequality holds uniformly over possible score realizations, and averaging over internal randomness preserves the bound. Thus, privacy does not require calibrated estimator errors or a probability that the estimators identify private dependencies correctly.

Step 2: Privacy of routing decisions. The operational decision $\widehat{z}_i = \mathcal{R}(\widehat{\mathrm{PD}}_i, \widehat{\mathrm{EE}}_i)$ is a deterministic function of the protected scores. Consequently, it is post-processing and adds no privacy loss beyond $\epsilon_i^{\mathrm{p}} + \epsilon_i^{\mathrm{e}}$. Whether a cloud request occurs is likewise covered, provided it follows this protected decision and is not subsequently overridden by an unprotected private-dependent condition.

For analysis, one may expose both protected scores in an augmented transcript. Bounding the privacy of this stronger observation also bounds the actual transcript in which the scores remain local.

Step 3: Privacy of message releases. Fix a public history t and its candidate set $\mathcal{C}_i(A, t)$. Write

$$Z_i(x, t) = \sum_{c \in \mathcal{C}_i(A,t)} \exp\left(\epsilon_i^{\mathrm{m}} v_i(x, t, c)/2\right). \tag{39}$$

The bounded score range gives

$$|v_i(x, t, c) - v_i(x', t, c)| \leq 1 \tag{40}$$

for every candidate. It also implies

$$Z_i(x', t) \leq e^{\epsilon_i^{\mathrm{m}}/2} Z_i(x, t). \tag{41}$$

Hence,

$$\begin{aligned} &\frac{\Pr(M_i = c \mid x, A, t)}{\Pr(M_i = c \mid x', A, t)} \\ &\quad = \exp\left(\frac{\epsilon_i^{\mathrm{m}}}{2}[v_i(x, t, c) - v_i(x', t, c)]\right) \frac{Z_i(x', t)}{Z_i(x, t)} \\ &\quad \leq e^{\epsilon_i^{\mathrm{m}}}. \end{aligned} \tag{42}$$

This step requires a common candidate set for the compared private inputs. It does not require the relevance score to be accurate.

The same proof establishes budget ϵ_0 for the initial query release. A message constructed solely from public information and preceding protected releases is post-processing and has zero additional privacy cost.

Step 4: Adaptive composition. Consider the augmented transcript that includes the protected score releases. At any fixed prefix, the next protected score pair has privacy loss at most $\epsilon_i^{\mathrm{p}} + \epsilon_i^{\mathrm{e}}$. A subsequent private-dependent message has conditional privacy loss at most ϵ_i^{m}. A local branch with no message has zero additional message-release cost.

Cloud responses, the public SQ-DAG, and other observable post-processing depend only on the cloud's existing information, protected releases, and cloud randomness. They add no further privacy

loss. The metadata restrictions ensure that the transcript contains no unaccounted private-dependent observations.

Multiplying the conditional likelihood-ratio bounds along a transcript of at most N routing steps gives

$$\frac{dP_x}{dP_{x'}}(Y) \le \exp\left(\epsilon_0 + \sum_{i=1}^{N}[\epsilon_i^{\mathrm{p}} + \epsilon_i^{\mathrm{e}} + \epsilon_i^{\mathrm{m}}]\right), \tag{43}$$

where $P_x = \mathcal{L}(Y \mid X = x, A = a)$. Removing the additional protected scores from the augmented transcript is post-processing. Integrating over any measurable event proves Eq. (20). Exchanging x and x' yields the reverse likelihood-ratio bound.

Step 5: A likelihood-ratio lemma. Suppose distributions P, Q satisfy

$$e^{-\epsilon} \le \frac{dP}{dQ} \le e^{\epsilon}. \tag{44}$$

Then

$$D_{\mathrm{KL}}(P\|Q) \le \epsilon \tanh(\epsilon/2). \tag{45}$$

For $\epsilon > 0$, write $r = dP/dQ$, so that $\mathbb{E}_Q[r] = 1$. Convexity of $r \log r$ bounds it above by the chord joining its values at $e^{-\epsilon}$ and e^{ϵ}. Taking expectations and using the mean constraint gives

$$\begin{aligned} D_{\mathrm{KL}}(P\|Q) &= \mathbb{E}_Q[r \log r] \\ &\le \epsilon \frac{e^{\epsilon} - 1}{e^{\epsilon} + 1} \\ &= \epsilon \tanh(\epsilon/2). \end{aligned} \tag{46}$$

The case $\epsilon = 0$ follows from $P = Q$.

Step 6: Mutual-information bounds. Fix $A = a$. Because Eq. (20) holds for every pair of compatible private inputs, the same likelihood-ratio bound holds between mixtures over those inputs. In particular, it holds between $P_{Y|S=s,A=a}$ and $P_{Y|A=a}$.

Applying Eq. (45),

$$\begin{aligned} I(S; Y \mid A = a) &= \sum_s \Pr(S = s \mid A = a)\, D_{\mathrm{KL}}\left(P_{Y|S=s,A=a} \middle\| P_{Y|A=a}\right) \\ &\le c(\epsilon_{\mathrm{tot}}). \end{aligned} \tag{47}$$

Averaging over A and using $I(S; Y \mid A) \le H(S \mid A)$ proves Eq. (22). Since A is part of Y, the chain rule gives

$$I(S; Y) = I(S; A) + I(S; Y \mid A), \tag{48}$$

which proves Eq. (23).

Step 7: AUC bound. Fix $A = a$ and define

$$P_b = \mathcal{L}(Y \mid S = b, A = a), \qquad b \in \{0, 1\}. \tag{49}$$

Both classes are assumed to have positive conditional probability. The mixture argument gives

$$P_1(\mathcal{B}) \le e^{\epsilon_{\mathrm{tot}}} P_0(\mathcal{B}), \qquad P_0(\mathcal{B}) \le e^{\epsilon_{\mathrm{tot}}} P_1(\mathcal{B}) \tag{50}$$

for every measurable event $\mathcal{B}$.

For any test with false-positive rate u and true-positive rate v, the first inequality implies

$$v \le e^{\epsilon_{\mathrm{tot}}} u. \tag{51}$$

Applying the second inequality to the complementary event gives

$$1 - u \le e^{\epsilon_{\mathrm{tot}}}(1 - v), \tag{52}$$

or equivalently,

$$v \le 1 - e^{-\epsilon_{\mathrm{tot}}}(1 - u). \tag{53}$$

These bounds also apply to randomized tests and therefore to the optimal ROC curve:

$$v(u) \le \min\left\{e^{\epsilon_{\mathrm{tot}}}u,\, 1 - e^{-\epsilon_{\mathrm{tot}}}(1-u)\right\}. \tag{54}$$

Integrating over $u \in [0,1]$ yields

$$\mathrm{AUC}_a^{\star}(Y) \le \int_0^1 \min\left\{e^{\epsilon_{\mathrm{tot}}}u,\, 1 - e^{-\epsilon_{\mathrm{tot}}}(1-u)\right\} du = \frac{e^{\epsilon_{\mathrm{tot}}}}{1+e^{\epsilon_{\mathrm{tot}}}}. \tag{55}$$

If $A \perp S$, both classes have the same distribution of public backgrounds. Integrating the conditional likelihood-ratio inequalities over this common distribution establishes the corresponding bound for the joint observation $Y = (A, T)$ and hence for the unconditional optimal AUC.

B.7 Scope and Experimental Interpretation

The guarantee concerns the complete mediated transcript, not only the text of individual messages. Its validity requires the protected interfaces, common candidate supports, budget limits, and metadata restrictions to be enforced. Keeping raw private documents local or applying heuristic desensitization alone is insufficient.

The MI bound limits additional information beyond the pre-existing public background A. It does not imply that $I(S;A)$ is small. Likewise, the conditional AUC bound must not be interpreted as an unconditional bound when A is correlated with S.

Finite-sample attack AUC is not the same quantity as optimal population AUC. Reported MI estimates must specify their estimation procedure, units, and any normalization. An attacker-based lower-bound estimate is not a certified upper bound on mutual information. Stage-wise MI measurements also cannot generally be summed to obtain joint transcript MI.

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 SPEAR with seven representative baselines that cover four edge-cloud deployment paradigms, as summarized in Table 12. The Cloud-Only group includes a fully cloud-based setting, which serves as a privacy-unconstrained quality upper bound, and RemoteRAG, which improves query privacy through embedding perturbation and restricted retrieval while still relying mainly on cloud-side RAG. The edge-cloud group includes MINIONS, CoGenesis, LSRP and PRISM, which represent different forms of edge-cloud collaboration, including selective context transmission, sketch- or logits-level guidance, privacy-preserving cloud guidance for edge-side execution, and sensitivity-aware query routing with local refinement. The Edge-Only baseline executes all queries locally with an edge-side SLM, providing the strongest privacy protection but limited answer quality due to constrained local model capacity. Together, these baselines allow us to evaluate whether SPEAR can achieve a better privacy-utility trade-off than fully cloud-based, fully local, and existing edge-cloud collaborative methods.

Table 12: Baselines spanning cloud-only, edge-cloud collaborative, and edge-only settings.

Type	Baseline	Description
Cloud-Only	Cloud-Only	All queries are handled by a cloud-side LLM with full access to user data, serving as a privacy-unconstrained upper bound.
	RemoteRAG	A cloud RAG framework close to the Cloud-Only setting, while using embedding perturbation and restricted retrieval to protect user-query privacy.
Edge-Cloud	MINIONS	A edge-cloud communication protocol that reduces cloud LLM costs on data-intensive reasoning tasks through selective context transmission from the edge to the cloud.
	CoGenesis	A edge-cloud collaborative generation method where the cloud LLM provides sketch- or logits-level guidance for edge execution.
	LSRP	A edge-cloud collaborative generation framework where the cloud LLM provides guidance for the edge-side SLM while private data remain on the edge device.
	PRISM	A privacy-aware routing method that routes queries by sensitivity, applies adaptive local differential privacy, and refines cloud-generated semantic sketches with edge context.
Edge-Only	Edge-Only	All queries are handled locally by an edge-side SLM, providing the strongest privacy protection but limited model capacity.

C.3 Evaluation Metrics

We evaluate SPEAR from two perspectives: answer quality and privacy leakage, as summarized in Table 13. The higher *F1*, *SM*, and *LaJ* are better, while *AUC* and *MI* are opposite.

Table 13: Evaluation metrics from the perspectives of answer quality and privacy leakage.

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 semantic agreement between the prediction and reference.
	LLM-as-a-Judge (LaJ) (Zheng et al., 2023)	Uses a strong LLM to assess answer correctness and relevance.
Privacy Leakage	Attribute Inference AUC (Mireshghallah et al., 2022; Wang et al., 2025)	Measures whether sensitive attributes can be inferred from transmission contexts.
	Mutual Information (MI) (Duan et al., 2024; Noorbakhsh et al., 2024)	Measures the information about sensitive attributes involved in transmission contexts.

Table 14: Cloud-side LLM and edge-side SLM settings.

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	Gemma3-1B	Mobile
LLM4	Claude-Sonnet-4.6	Closed-Source	SLM4	GLM-Edge-1.5B	Mobile

C.4 LLM and SLM Settings

We evaluate SPEAR with multiple cloud-side LLMs and edge-side SLMs settings as shown in Table 14, spanning different model families, capacities, sources, and deployment settings (Bai et al., 2023; Achiam et al., 2023; Team et al., 2023; Touvron et al., 2023; Bai et al., 2022; Glm et al., 2024).

C.5 Implementation Details

Our experimental edge devices fall into two categories: desktop computers and mobile devices.

Desktop Environment. For desktop settings, we consider three representative GPU configurations: NVIDIA GeForce RTX 3080, RTX 2080 Ti, and RTX 2080. These devices cover different levels of desktop edge capability and are used to evaluate how SPEAR behaves under varying local computation resources. By comparing these configurations, we can analyze the impact of edge-side hardware capacity on latency, local subquery execution, and overall system efficiency.

Mobile Deployment. Beyond desktop settings, we further evaluate SPEAR on a real-world mobile device to assess its practicality under resource-constrained edge environments. The mobile device used in our experiments is OPPO Find X8, a common smartphone with moderate on-device computing capability. On the OPPO Find X8 device, models are deployed via a lightweight inference pipeline built on Android Studio and ONNX Runtime Mobile. This setup provides a testbed for examining the deployability and efficiency of SPEAR on consumer-grade mobile hardware.

Software Environment. All desktop experiments are conducted using Python 3.10 with PyTorch 2.1 and CUDA 12.1. We adopt the Hugging Face Transformers (v4.40) and Accelerate frameworks to manage model loading, inference, and device placement. For efficient local inference on edge GPUs, we enable mixed-precision computation (FP16) and utilize FlashAttention where supported. On mobile devices, we deploy models using ONNX Runtime, with additional optimization via quantization (INT8) to meet memory and latency constraints. We deploy the open-source ExecuteTorch framework for mobile execution (GitHub repository: `https://github.com/pytorch/executorch`).

For local execution, we employ distilled or smaller-scale models (e.g., GLM-Edge-1.5B) hosted directly on edge devices. For cloud-side inference, we invoke proprietary or open-weight LLM APIs.

Containerization and Simulation Environment. To ensure consistent runtime behavior, all desktop experiments are encapsulated within Docker containers based on NVIDIA CUDA runtime images. Each container includes pre-installed dependencies, GPU drivers, and model checkpoints.

Communication. Edge-cloud communication is implemented via HTTP/2-based APIs with JSON serialization.

D Overall Performance on Other LLM+SLM Settings

Table 15 reports additional overall performance results under LLM3/LLM4 and SLM3/SLM4 settings. The results are consistent with previous observations: SPEAR consistently achieves the best privacy-utility trade-off among all privacy-preserving methods, ranking close to Cloud-Only in answer quality while maintaining much lower privacy leakage. Across all three datasets, SPEAR outperforms RemoteRAG, MINIONS, CoGenesis, and LSRP in F1, SM, and LaJ in most settings, and keeps AUC and MI close to Edge-Only. Compared with the main results using LLM1/LLM2 and SLM1/SLM2, these additional settings involve stronger closed-source cloud LLMs and more resource-constrained mobile SLMs, leading to slightly larger variations in exact answer quality and privacy leakage.

Nevertheless, the relative ordering remains stable: Cloud-Only gives the quality upper bound, Edge-Only gives the privacy upper bound, and SPEAR provides the most balanced solution.

Table 15: Overall performance comparison between SPEAR, Cloud/Edge-Only and SOTA methods.

Method	P-PubMedQA					P-FinQA					P-MultiHopRAG				
	Accuracy			Privacy		Accuracy			Privacy		Accuracy			Privacy	
	F1↑	SM↑	LaJ↑	AUC↓	MI↓	F1↑	SM↑	LaJ↑	AUC↓	MI↓	F1↑	SM↑	LaJ↑	AUC↓	MI↓
LLM3+SLM3															
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
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	.467	.524	.573	**.500**	**.000**	.534	.605	.562	**.500**	**.000**	.463	.569	.536	**.500**	**.000**
SPEAR (*Ours*)	.662	.717	.712	.541	.121	.738	.756	.761	.543	.161	.594	.729	.732	.547	.168
LLM3+SLM4															
Cloud-Only	**.665**	**.682**	**.793**	1.0	1.0	**.648**	**.764**	**.773**	1.0	1.0	**.569**	**.763**	**.772**	1.0	1.0
RemoteRAG	.613	.638	.729	.672	.208	.597	.684	.722	.659	.188	.525	.706	.724	.658	.195
MINIONS	.583	.617	.712	.655	.183	.570	.639	.677	.643	.159	.496	.679	.673	.641	.174
CoGenesis	.566	.596	.652	.622	.160	.542	.583	.635	.622	.148	.478	.647	.648	.628	.163
LSRP	.519	.561	.620	.596	.153	.511	.546	.593	.593	.132	.455	.596	.614	.601	.147
PRISM	.605	.619	.718	.585	.146	.581	.667	.697	.573	.117	.502	.683	.711	.585	.138
Edge-Only	.493	.532	.576	**.500**	**.000**	.497	.524	.561	**.500**	**.000**	.443	.542	.563	**.500**	**.000**
SPEAR (*Ours*)	.636	.669	.773	.571	.139	.629	.747	.754	.561	.106	.536	.739	.758	.569	.131
LLM4+SLM3															
Cloud-Only	**.687**	**.719**	**.796**	1.0	1.0	**.744**	**.773**	**.785**	1.0	1.0	**.577**	**.793**	**.796**	1.0	1.0
RemoteRAG	.624	.659	.718	.676	.188	.683	.712	.719	.673	.226	.535	.698	.707	.669	.231
MINIONS	.597	.624	.685	.655	.172	.664	.693	.692	.662	.207	.517	.672	.683	.654	.214
CoGenesis	.569	.601	.648	.646	.165	.641	.667	.673	.652	.193	.502	.658	.647	.639	.198
LSRP	.548	.583	.625	.592	.143	.628	.632	.661	.603	.178	.482	.611	.629	.587	.167
PRISM	.617	.634	.702	.568	.139	.669	.704	.675	.588	.165	.532	.687	.693	.557	.161
Edge-Only	.511	.543	.604	**.500**	**.000**	.592	.599	.618	**.500**	**.000**	.462	.587	.593	**.500**	**.000**
SPEAR (*Ours*)	.668	.696	.778	.552	.128	.718	.751	.765	.566	.159	.553	.747	.751	.548	.154
LLM4+SLM4															
Cloud-Only	**.659**	**.697**	**.793**	1.0	1.0	**.725**	**.798**	**.786**	1.0	1.0	**.575**	**.758**	**.789**	1.0	1.0
RemoteRAG	.620	.642	.722	.645	.193	.674	.746	.739	.657	.204	.523	.706	.727	.649	.211
MINIONS	.594	.617	.703	.613	.169	.649	.712	.705	.634	.183	.493	.668	.683	.625	.197
CoGenesis	.577	.598	.679	.596	.155	.613	.682	.657	.611	.166	.472	.647	.658	.596	.188
LSRP	.561	.576	.647	.589	.147	.595	.646	.629	.593	.148	.461	.619	.632	.578	.182
PRISM	.598	.627	.709	.567	.122	.654	.729	.717	.573	.134	.504	.682	.691	.572	.176
Edge-Only	.536	.544	.609	**.500**	**.000**	.562	.608	.594	**.500**	**.000**	.438	.587	.601	**.500**	**.000**
SPEAR (*Ours*)	.642	.671	.776	.558	.109	.699	.781	.768	.544	.119	.550	.733	.764	.551	.165

Table 16: Privacy leakage analysis of SPEAR across four transmission stages.

Metric	Model	P-PubMedQA				P-FinQA				P-MultiHopRAG			
		q2c	sq2e	pq2c	pa2e	q2c	sq2e	pq2c	pa2e	q2c	sq2e	pq2c	pa2e
AUC↓	LLM3+SLM3	.563	.535	.534	.532	.559	.541	.531	.541	.564	.543	.548	.533
	LLM3+SLM4	.586	.568	.559	.571	.578	.562	.553	.551	.583	.561	.560	.572
	LLM4+SLM3	.558	.546	.554	.550	.572	.564	.567	.561	.557	.543	.539	.553
	LLM4+SLM4	.569	.555	.551	.557	.553	.542	.538	.543	.558	.542	.549	.555
MI↓	LLM3+SLM3	.038	.027	.031	.025	.049	.037	.035	.040	.052	.039	.042	.035
	LLM3+SLM4	.042	.036	.030	.031	.033	.025	.027	.021	.042	.034	.029	.026
	LLM4+SLM3	.038	.028	.031	.031	.052	.043	.033	.031	.047	.035	.039	.033
	LLM4+SLM4	.029	.023	.029	.028	.035	.031	.028	.025	.055	.038	.041	.031

E Further Analysis of Privacy Leakage

E.1 Privacy Leakage on Other LLM+SLM Settings

Table 16 reports additional stage-wise privacy leakage results under LLM3+4 and SLM3+4 settings. The results are broadly consistent with LLM1+2 and SLM1+2 settings: the total AUC remains

close to random guessing, and MI stays low across all datasets, indicating that SPEAR still exposes only limited sensitive signals under additional model combinations. The stage-wise pattern is also similar, with edge-to-cloud transmissions generally carrying slightly higher leakage than cloud-to-edge outputs. Compared with previous results, these settings involve stronger closed-source cloud LLMs and more resource-constrained mobile SLMs, leading to slightly larger AUC variation but comparable or even lower MI in some cases. Overall, the results further confirm that SPEAR's privacy behavior remains stable across diverse edge-cloud model settings.

E.2 Privacy Leakage under Different Attacker Strengths

Table 17 evaluates privacy leakage under attackers with different strengths. As expected, stronger attackers lead to higher AUC and MI across all three datasets, with the LLM-based attacker producing the largest leakage. Nevertheless, even under the strongest attacker, the AUC remains close to random guessing, ranging from 0.533 to 0.546, and the MI remains limited. This suggests that **SPEAR does not rely on a weak attacker assumption**; instead, its transmitted contents contain only limited sensitive signals even when analyzed by stronger inference models. The leakage increase is relatively more visible on P-FinQA and P-MultiHopRAG than on P-PubMedQA, indicating that financial and multi-hop tasks may expose richer indirect cues.

Table 17: Privacy leakage under different strengthes of attackers.

Attacker Strength	Description	P-PubMedQA		P-FinQA		P-MultiHopRAG	
		AUC	MI	AUC	MI	AUC	MI
Weak	Bag-of-Words Attacker	.514	.067	.517	.109	.521	.126
Medium	Encoder-Based Attacker	.521	.085	.528	.127	.529	.158
Strong	LLM-Based Attacker	.533	.116	.538	.164	.546	.175

E.3 Content Analysis of Privacy Leakage

Table 18 analyzes which types of private information are most vulnerable to leakage. Behavioral information ranks first, with the highest AUC and MI, suggesting that user behavior patterns may leave more persistent cues in transmitted contexts. Internal decision information and medical history also show relatively higher leakage, indicating that decision constraints and longitudinal records are harder to fully remove during desensitization and subquery transformation. In contrast, diagnosis results and debt records exhibit lower leakage among the top categories, implying that more explicit private attributes can be more effectively localized or filtered by SPEAR. Overall, the results show that **residual leakage mainly comes from implicit contextual cues rather than directly exposed private fields**.

Table 18: Top-5 categories of privacy leakage.

Top-K	Categories	AUC	MI
1	Behavioral Information	.554	.186
2	Internal Decision	.549	.155
3	Medical History	.541	.131
4	Diagnosis Result	.534	.117
5	Debt Record	.532	.112

F Ablation Study on Other Datasets

F.1 Ablation Study of PDE, EEE and ECR

Table 19 provides additional ablation results on P-FinQA and P-MultiHopRAG. Consistent with the main P-PubMedQA results, the full SPEAR model achieves the best privacy-utility trade-off, while removing any of PDE, EEE, or ECR degrades both answer quality and privacy protection. PDE remains the key component for privacy control, as its removal consistently increases AUC and MI. EEE improves the reliability of edge execution, and ECR is essential for translating privacy and

executability estimates into effective routing decisions. Notably, keeping only ECR without PDE and EEE again leads to severe performance collapse and high privacy leakage, confirming that routing must be supported by accurate estimation signals. Compared with P-PubMedQA, P-FinQA and P-MultiHopRAG show more evident privacy degradation after ablation, indicating that financial and multi-hop tasks are more sensitive to incorrect routing. These results further validate the generality of SPEAR across domains.

Table 19: The ablation of modules.

PDE	EEE	ECR	P-FinQA					P-MultiHopRAG				
			Accuracy			Privacy		Accuracy			Privacy	
			F1	SM	LaJ	AUC	MI	F1	SM	LaJ	AUC	MI
✓	✓	✓	.734	.739	.763	.538	.164	.587	.767	.724	.546	.175
✗	✓	✓	.669	.687	.704	.652	.224	.522	.713	.665	.637	.216
✓	✗	✓	.647	.652	.662	.591	.186	.513	.685	.649	.603	.197
✓	✓	✗	.603	.621	.637	.644	.208	.501	.652	.617	.631	.203
✓	✗	✗	.558	.589	.603	.632	.196	.467	.601	.578	.612	.199
✗	✓	✗	.531	.563	.587	.668	.239	.443	.557	.542	.652	.227
✗	✗	✓	.274	.341	.387	.906	.486	.256	.316	.345	.913	.478

F.2 Ablation Study of PDE Indicators

Table 20 further validates the complementary roles of G_i^{pri} and D_i^{slot} on P-FinQA and P-MultiHopRAG. As in P-PubMedQA, the full PDE achieves the best privacy-utility trade-off, while removing either indicator degrades answer quality and increases leakage. Removing D_i^{slot} causes larger privacy degradation, highlighting the importance of structural private-slot estimation, whereas G_i^{pri} provides complementary functional evidence. Compared with P-PubMedQA, the effect of D_i^{slot} is more pronounced on P-FinQA and P-MultiHopRAG, suggesting that financial and multi-hop tasks are more sensitive to structural private slots. The worst results occur when both indicators are removed, confirming the necessity of combining both signals.

Table 20: The ablation of PDE indicators.

G_i^{pri}	D_i^{slot}	P-FinQA					P-MultiHopRAG				
		Accuracy			Privacy		Accuracy			Privacy	
		F1	SM	LaJ	AUC	MI	F1	SM	LaJ	AUC	MI
✓	✓	.734	.739	.763	.538	.164	.587	.767	.724	.546	.175
✓	✗	.678	.694	.715	.598	.198	.537	.646	.672	.587	.211
✗	✓	.685	.698	.724	.561	.172	.544	.692	.688	.572	.187
✗	✗	.669	.687	.704	.652	.224	.522	.713	.665	.637	.216

F.3 Ablation Study of EEE Indicators

Table 21 examines the two EEE signals on P-FinQA and P-MultiHopRAG. The complete EEE consistently performs best, showing that executability estimation benefits from both answer quality and stability. Removing D_i^{qual} causes a larger drop than removing D_i^{cons}, suggesting that answer quality is the leading signal, while self-consistency mainly provides stabilizing evidence. Unlike P-PubMedQA, the degradation is more evident on the financial and multi-hop datasets, indicating that these tasks are more sensitive to inaccurate executability estimation. The worst results occur when both signals are removed, confirming their complementarity.

F.4 Ablation Study of Routing Strategies in ECR

Table 22 highlights the different roles of the two routing mechanisms in ECR. The hard rule mainly serves as a safety filter for clearly private-dependent subqueries, while the soft score provides the flexibility needed for ambiguous cases. This distinction is especially important on P-FinQA and P-MultiHopRAG, where many subqueries involve mixed privacy and capability requirements. Compared with P-PubMedQA, the benefit of the soft score is more pronounced, indicating that

Table 21: The ablation of EEE indicators.

D_i^{qual}	D_i^{cons}	P-FinQA					P-MultiHopRAG				
		Accuracy			Privacy		Accuracy			Privacy	
		F1	SM	LaJ	AUC	MI	F1	SM	LaJ	AUC	MI
✓	✓	.734	.739	.763	.538	.164	.587	.767	.724	.546	.175
✓	✗	.682	.691	.704	.567	.179	.543	.702	.673	.577	.187
✗	✓	.658	.677	.682	.583	.182	.516	.693	.659	.592	.196
✗	✗	.647	.652	.662	.591	.186	.513	.685	.649	.603	.197

financial and multi-hop reasoning requires more adaptive routing. Removing both mechanisms leads to the weakest results, showing that effective ECR depends on both privacy-first filtering and flexible score-based routing.

Table 22: The ablation of ECR.

z_i Strategy		P-FinQA					P-MultiHopRAG				
		Accuracy			Privacy		Accuracy			Privacy	
Hard	Soft	F1	SM	LaJ	AUC	MI	F1	SM	LaJ	AUC	MI
✓	✓	.734	.739	.763	.538	.164	.587	.767	.724	.546	.175
✓	✗	.652	.664	.685	.581	.194	.533	.676	.641	.598	.194
✗	✓	.677	.691	.724	.548	.173	.562	.712	.686	.557	.181
✗	✗	.603	.621	.637	.644	.208	.501	.652	.617	.631	.203

G Overhead Analysis on Mobile Edge Settings

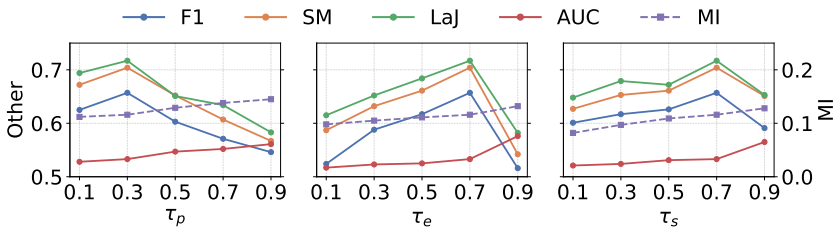
Table 23 reports the stage-wise latency of SPEAR under mobile edge (OPPO Find X8) and SLM settings. Compared with desktop edge devices in Table 10, **mobile settings lead to higher total latency**, mainly due to increased costs in desensitization, PDE, EEE, ECR, fusion, and final answer generation. Nevertheless, the overall latency remains within a practical range, and the dominant cost still comes from cloud-side decomposition and edge-side subquery execution. SLM3 achieves lower latency than SLM4 across most stages, indicating that mobile model choice also affects system efficiency.

Table 23: Stage-wise latency of SPEAR under mobile edge SLM settings, measured in seconds.

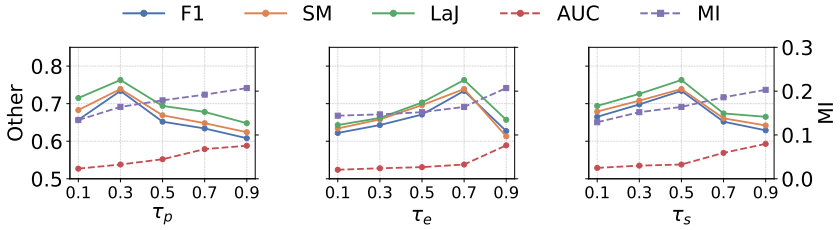
Model	Stage										Total
	Desen.	Q2Cloud	Decom.	SQ2Edge	PDE	EEE	ECR	SQExecution	Fusion	Answer	
SLM3	4.28	2.07	45.82	2.19	2.23	1.64	2.12	13.06	5.58	2.44	82.43
SLM4	5.52	2.65	46.59	2.77	2.89	2.27	2.74	14.61	6.94	3.18	89.16

H Sensitivity Analysis of Hyperparameters on Other Datasets

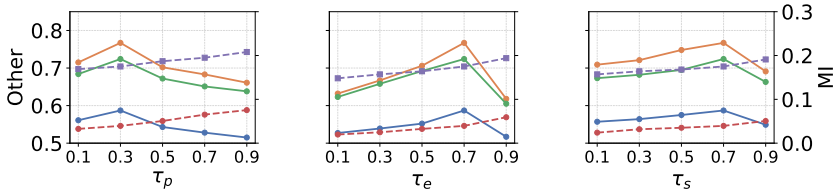
Fig. 3 summarizes hyperparameter sensitivity across P-PubMedQA, P-FinQA, and P-MultiHopRAG. SPEAR remains broadly stable across a wide range of settings, with consistent design preferences across datasets. For the routing thresholds, τ_p generally favors low-to-moderate values, whereas τ_e and τ_s tend to favor higher values. For the weighting coefficients, the preferred α suggests slightly greater emphasis on structural privacy cues in P-PubMedQA and a more balanced combination of privacy signals in P-FinQA and P-MultiHopRAG. Larger γ values favor answer quality over consistency in EEE, while a moderately privacy-aware routing weight η provides a favorable privacy–utility trade-off in ECR. Despite these shared trends, τ_s and η exhibit more pronounced task-dependent variation on P-FinQA and P-MultiHopRAG, indicating that their preferred settings depend on the dataset. Among the evaluated hyperparameters, η is the most sensitive, highlighting the importance of balancing privacy dependency and edge executability in the final routing decision.



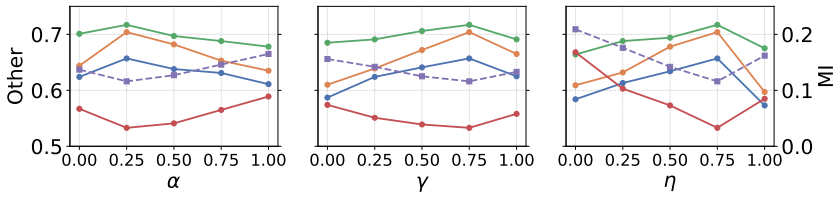
(a) Threshold sensitivity on P-PubMedQA.



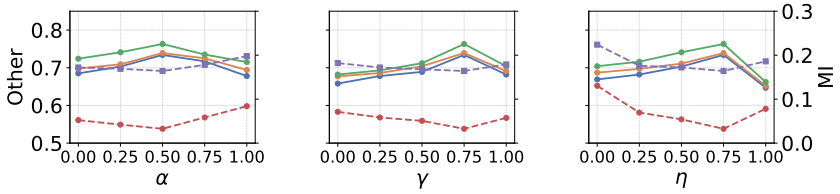
(b) Threshold sensitivity on P-FinQA.



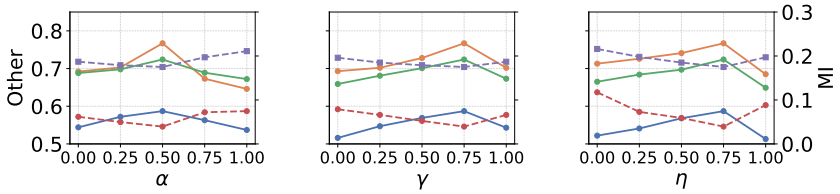
(c) Threshold sensitivity on P-MultiHopRAG.



(d) Weight sensitivity on P-PubMedQA.



(e) Weight sensitivity on P-FinQA.



(f) Weight sensitivity on P-MultiHopRAG.

Figure 3: Hyperparameter sensitivity across three datasets. Panels (a)–(c) vary the routing thresholds τ_p, τ_e, and τ_s; panels (d)–(f) vary the weighting coefficients.

I Sensitivity Analysis of Function Implementations on Other Datasets

Table 24 shows that the function choices used in the main experiments generalize well to P-FinQA and P-MultiHopRAG. Specifically, KSS, SWS, TIC, and NG consistently achieve the best answer quality and lowest privacy leakage for $\mathrm{Dir}(\cdot)$, $\mathrm{Spec}(\cdot)$, $w(r)$, and $\mathrm{Cons}(\cdot)$, respectively. Compared with alternative implementations, these choices provide more accurate function estimation and lead to more reliable routing decisions, which in turn improves both utility and privacy. This is important because these functions directly affect PDE and EEE, and inaccurate estimates may cause privacy-dependent subqueries to be incorrectly offloaded or executable subqueries to be unnecessarily kept local. The advantage is particularly clear for $\mathrm{Cons}(\cdot)$, where NG outperforms semantic similarity and keyword coverage, suggesting that conclusion-level agreement is better captured by surface-form consistency in our setting. Overall, the results align with the P-PubMedQA observations and support the cross-domain robustness of the selected function implementations.

Table 24: The sensitivity of function implementations on P-FinQA and P-MultiHopRAG.

Function	Method	P-FinQA					P-MultiHopRAG				
		Accuracy			Privacy		Accuracy			Privacy	
		F1	SM	LaJ	AUC	MI	F1	SM	LaJ	AUC	MI
$\mathrm{Dir}(\cdot)$	KSS	.734	.739	.763	.538	.164	.587	.767	.724	.546	.175
	KLD	.721	.713	.711	.557	.188	.564	.732	.681	.574	.193
$\mathrm{Spec}(\cdot)$	SWS	.734	.739	.763	.538	.164	.587	.767	.724	.546	.175
	CSD	.719	.722	.748	.545	.169	.566	.752	.715	.553	.181
$w(r)$	TIC	.734	.739	.763	.538	.164	.587	.767	.724	.546	.175
	KC	.721	.716	.752	.549	.183	.559	.743	.706	.558	.189
$\mathrm{Cons}(\cdot)$	NG	.734	.739	.763	.538	.164	.587	.767	.724	.546	.175
	SS	.718	.712	.743	.553	.179	.565	.754	.706	.557	.193
	KC	.704	.693	.724	.563	.186	.541	.736	.682	.572	.201

J Case Study

This case study illustrates how SPEAR performs privacy-aware edge-cloud collaborative reasoning over a concrete medical query, where *gray* and *purple* boxes denote edge-side and cloud-side executions, respectively.

Edge Desensitization removes or generalizes identifiable and sensitive details, such as the patient name and the specific disease name, before the query is sent to the cloud, while preserving the original clinical intent. This ensures that the cloud can understand the task goal without directly observing the raw query or private medical attributes.

Cloud Decomposition then decomposes the desensitized medical recommendation task into a dependency-structured subquery sequence, enabling public and private reasoning steps to be handled separately. In this way, the cloud is used for high-level planning without accessing the local private knowledge base.

Edge-Cloud Routing estimates high privacy dependency and high edge executability for Subquery 3, leading ECR to route it to the edge rather than the cloud. This decision keeps private clinical records, such as spirometry, oxygen saturation, chest X-ray, and CBC results, local to the edge, while still allowing privacy-insensitive subqueries to benefit from cloud-side capability.

Edge Fusion finally combines public subanswers with private clinical evidence to generate the diagnostic recommendation.

Edge Desensitization

`Instruction:`

You are a privacy-preserving text rewriting assistant.

Your task is to rewrite the given query by removing or anonymizing any sensitive or identifiable information, while preserving the original intent and meaning.

`Raw Query:`

I'm planning to go for a follow-up visit. My medical history is chronic obstructive pulmonary disease (COPD). Recently, I have been engaged in prolonged heavy physical labor in a dust-exposed environment. Based on my prior medical records and current clinical findings, appropriate diagnostic and treatment recommendations are provided.

`Desensitized Query:`

I plan to have follow-up visit for my respiratory disorder. Recently, I have been engaged in prolonged heavy physical labor in a dust-exposed environment. Based on my prior medical records and current clinical findings, appropriate diagnostic and treatment recommendations will be made.

Cloud Decomposition

`Instruction:`

You are a query decomposition planner.

Break a complex question into multi-subqueries.

Ensure that the subquery does not trigger a refusal response.

`Question:`

I am here for a follow-up visit for my chronic respiratory condition. Recently, I have been engaged in prolonged heavy physical labor in a dust-exposed environment. Based on my prior medical records and current clinical findings, appropriate diagnostic and treatment recommendations will be made.

`Subqueries in JSON:`

[

{"step": 1, "sub_query": "What baseline clinical factors and symptoms should be assessed during my follow-up visits for a chronic respiratory condition?", "depends_on": null},

{"step": 2, "sub_query": "Given these baseline assessments, how does prolonged heavy physical labor in a dust-exposed environment modify my risk or symptom profile?", "depends_on": 1},

{"step": 3, "sub_query": "Based on my updated risk and symptom profile, what diagnostic evaluations should be prioritized?", "depends_on": 2}

]

Edge-Cloud Routing Subquery 3

`Subquery:`

Based on my updated risk and symptom profile, what diagnostic evaluations should be prioritized?

`Indicators:`

PDE: $\mathrm{PD}_i = \alpha G_i^{pri} + (1-\alpha) D_i^{slot} \rightarrow 0.25 \times 0.672 + 0.75 \times 0.746 = 0.728 \geq 0.3 \rightarrow yes$

EEE: $\mathrm{EE}_i = \gamma D_i^{qual} + (1-\gamma) D_i^{cons} \rightarrow 0.75 \times 0.843 + 0.25 \times 0.691 = 0.805 \geq 0.7 \rightarrow yes$

ECR: $S_i = \eta\, \mathrm{PD}_i + (1-\eta)\, \mathrm{EE}_i \rightarrow 0.75 \times 0.728 + 0.25 \times 0.805 = 0.747 \geq 0.7 \rightarrow yes$

`Routing Decision:`

Subquery 3 should be edge-routed due to its strong private dependency and sufficient edge executability.

Edge Fusion

`Instruction:`

You are an answer synthesis assistant.

Your task is to combine multiple subanswers into a single coherent final answer.

`Subanswers:`

Follow-up assessment includes symptoms, lung function, exacerbation history, and treatment adherence.

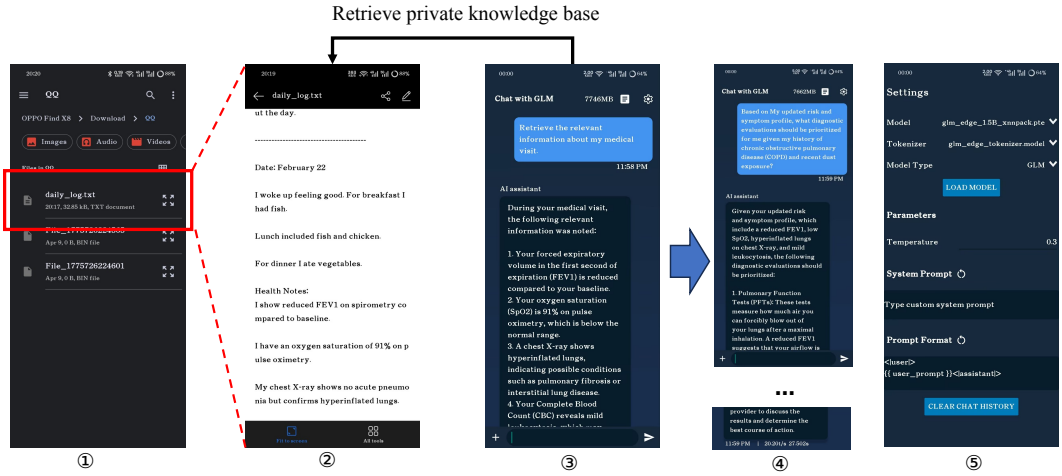
Dust exposure and heavy labor increase respiratory irritation, worsen symptoms, and raise exacerbation risk.

I shows reduced FEV1 on spirometry compared to baseline. I have an oxygen saturation of 91% on pulse oximetry. My chest X-ray shows no acute pneumonia but confirms hyperinflated lungs. My CBC indicates mild leukocytosis.

`Final Answer:`

Priority evaluations include arterial blood gas analysis to assess hypoxemia and CO_2 retention, and sputum testing to rule out infection. Repeat spirometry and close monitoring are also recommended to track disease progression.

K Real-World Mobile Device Demonstration



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

We further deploy a SPEAR-enhanced RAG system on a real-world smartphone, OPPO Find X8, to demonstrate the practical applicability of SPEAR in mobile edge scenarios.

The OPPO Find X8 is equipped with a MediaTek Dimensity 9400 SoC, which provides heterogeneous computing resources, including multi-core Armv9 CPUs, an integrated GPU, and a dedicated neural processing unit (NPU) for on-device AI acceleration. With up to 12 GB RAM and UFS 4.0 storage, the device can support lightweight edge-side language model inference. In our deployment, we use GLM-Edge-1.5B as the edge-side SLM on the smartphone, as seen in Fig. 4 ⑤.

As shown in Fig. 4 ①, the private knowledge base is stored locally on the Android file system of OPPO Find X8. The device can retrieve private daily records from the local file, such as personal health notes, medical measurements, and clinical observations in Fig. 4 ②③, without uploading the raw private knowledge to the cloud. This setting follows the privacy constraint of SPEAR, where private knowledge remains on the edge throughout the collaborative generation process.

In this mobile SPEAR system, as shown in Fig. 4 ④, the smartphone receives the user query, retrieves relevant private medical knowledge, performs privacy-sensitive local processing, and generates the final response for the user. Private-dependent subqueries are executed locally by GLM-Edge-1.5B, while computationally intensive and privacy-insensitive public subqueries are offloaded to the cloud-side LLM, Qwen3.5-397B. The cloud-returned public answers are then sent back to the smartphone and fused with local private reasoning results for final answer generation.

This real-device demonstration shows that SPEAR can be practically deployed on mobile edge devices, effectively mitigating the capability and privacy mismatch between edge-side SLMs and cloud-side LLMs in privacy-sensitive RAG scenarios.

L LIMITATIONS

While SPEAR effectively enables fine-grained privacy- and executability-aware edge-cloud collaboration for multi-step RAG tasks, several aspects merit further exploration. Currently, privacy dependency and edge executability are estimated using fixed heuristics and static models; the impact of alternative estimation strategies or adaptive learning on routing performance remains underexplored. Moreover, the quality of subquery decomposition and dependency graph is ultimately limited by the capability of the underlying models. In future work, we plan to investigate adaptive estimators and dynamic strategies to further enhance the flexibility, robustness, and overall quality of collaborative generation under privacy constraints.