

# Beyond Local vs. External: A Game-Theoretic Framework for Trustworthy Knowledge Acquisition

Anonymous ACL submission

## Abstract

Cloud-hosted Large Language Models (LLMs) offer unmatched reasoning capabilities and dynamic knowledge, yet submitting raw queries to these external services risks exposing sensitive user intent. Conversely, relying exclusively on trusted local models preserves privacy but often compromises answer quality due to limited parameter scale and knowledge. To resolve this dilemma, we propose Game-theoretic Trustworthy Knowledge Acquisition (GTKA), a framework that formulates the trade-off between knowledge utility and privacy as a strategic game. GTKA consists of three components: (i) a privacy-aware sub-query generator that decomposes sensitive intent into generalized, low-risk fragments; (ii) an adversarial reconstruction attacker that attempts to infer the original query from these fragments, providing adaptive leakage signals; and (iii) a trusted local integrator that synthesizes external responses within a secure boundary. By training the generator and attacker in an alternating adversarial manner, GTKA optimizes the sub-query generation policy to maximize knowledge acquisition accuracy while minimizing the reconstructability of the original sensitive intent. To validate our approach, we construct two sensitive-domain benchmarks in the biomedical and legal fields. Extensive experiments demonstrate that GTKA significantly reduces intent leakage compared to state-of-the-art baselines while maintaining high-fidelity answer quality.<sup>1</sup>

## 1 Introduction

Large Language Models (LLMs) have become foundational tools for knowledge acquisition (Oruganti et al., 2023; Pondel et al., 2024), enabling knowledge-intensive organizations, including research institutions, enterprises, and strategic teams, to navigate vast corpora and synthesize information with unprecedented speed (Wang et al., 2025;

<sup>1</sup>All data and code will be publicly available; an anonymized copy is attached with this submission for review.

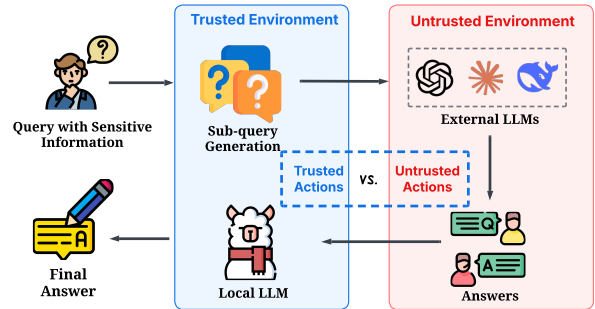


Figure 1: An illustration of our method.

Aamer et al., 2025; Lai et al., 2024; Lewis et al., 2020). By sharply reducing the time and cost of accessing domain knowledge, LLMs are now integral to daily research workflows (Spatharioti et al., 2025). However, this progress introduces a critical risk: many state-of-the-art LLMs are hosted on external cloud platforms (e.g., GPT-5), and submitting raw queries to these services can inadvertently expose high-value organizational intent (Su et al., 2024; Carlini et al., 2021). Such exposure can result in premature disclosure, idea appropriation, and competitive disadvantage, since intent-bearing queries may be logged, profiled, or inferred by external providers. For instance, consider a biomedical researcher investigating a novel therapeutic target for a specific cancer. A single query to a cloud-based LLM about highly specific compounds or mechanisms could expose their entire research trajectory. In these settings, intent leakage does not merely threaten individual privacy but can compromise organizational competitiveness and long-term value, long before any public disclosure occurs.

A large and growing literature seeks to curb information leakage from multiple angles (Sweeney, 2002). For example, query-side transformations, such as paraphrasing and semantic obfuscation, attempt to hide intent but create a blunt trade-off: aggressive transformations distort the user’s core information need, while subtle changes fail to meaningfully reduce risk (Jagwani and Kaushik, 2017).

072 Training-time protections, such as differential privacy, harden model parameters but leave the live  
073 query channel exposed (Abadi et al., 2016; Pang  
074 et al., 2024).  
075

076 Cryptographic and hardware-based solutions, including Multi-Party Computation (MPC), Homomorphic Encryption (HE), and Trusted Execution Environments (TEEs), offer significantly stronger confidentiality guarantees (Beimel, 2011; Chatopadhyay et al., 2024; Pillai and Polimetla, 2024; Gentry, 2009; Ohrimenko et al., 2016). However, unlike client-side semantic strategies, these methods require fundamental changes to the service hosting the model. Furthermore, MPC and HE suffer from prohibitive computational overheads. While TEEs provide a promising and efficient hardware-backed alternative, they necessitate migrating current deployments into encrypted virtual machines, representing a widespread infrastructure overhaul likely requiring years to mature (Mo et al., 2021). Moreover, even if the primary LLM is secured, downstream external tools may still fall outside the trusted perimeter.

095 To address these challenges, we propose a novel semantics-level framework that simultaneously leverages cloud-hosted LLMs for knowledge acquisition while safeguarding research intent from disclosure. As illustrated in Figure 1, given a sensitive user query, we first strategically decompose it into a set of semantically fragmented, low-risk subqueries. These subqueries are then dispatched to multiple, diverse cloud-hosted LLM endpoints to acquire the necessary external knowledge. Finally, a local LLM, which operates within the user’s trusted environment and is the only component to see the original sensitive context, aggregates the knowledge fragments returned from the external LLM services. A coherent, high-quality final answer is obtained. Compared with prior approaches, our method offers two key advantages. First, it leverages existing hosted LLMs without modification, thereby avoiding the complexity and inefficiency of cryptographic or infrastructure-heavy solutions. Second, it fundamentally enhances user trust and security by ensuring that raw queries containing sensitive intent are never directly exposed to any external, untrusted service.

119 The contributions are summarized as follows:

- We propose a game-theoretic framework for trustworthy knowledge acquisition (GTKA), which enables high-quality knowledge acquisition

123 from cloud-hosted LLMs while safeguarding sensitive user intent.  
124

- We cast the privacy–utility trade-off as an adversarial game between a privacy-aware sub-query generator and an adversarial reconstruction attacker. We optimize the sub-query policy via alternating adversarial training to jointly maximize answer quality and minimize intent reconstructability.  
125  
126  
127  
128  
129  
130  
131
- We construct two sensitive-domain benchmarks focused on biomedical and legal contexts. Extensive evaluations demonstrate that GTKA substantially reduces intent leakage against strong baselines while maintaining high-fidelity answer quality. To foster future research, we will release the benchmark, code, and models to the community<sup>2</sup>.  
132  
133  
134  
135  
136  
137  
138  
139

## 2 Related Work 140

### 2.1 Privacy Protection for Large Language Models 141 142

143 As large language models (LLMs) increasingly mediate human–AI knowledge exchange, concerns over privacy leakage have become central to their responsible deployment (Das et al., 2025; Chen et al., 2025). Prior research has explored diverse mechanisms to mitigate information exposure across the entire LLM lifecycle. Early work emphasizes parameter-level protections such as differential privacy and secure aggregation to prevent memorization of sensitive data during training (Abadi et al., 2016; Lyu et al., 2020). Subsequent studies extend protection to inference and deployment, employing confidential execution environments and encrypted federated pipelines to safeguard user data from server-side inspection (Yin et al., 2021; Zhao and Song, 2024). More recently, the focus has shifted toward the interaction layer, where user prompts themselves become a major source of leakage risk. Semantic rewriting and paraphrastic obfuscation (Jagwani and Kaushik, 2017), along with selective local sanitization (Kan et al., 2023), attempt to conceal intent before transmission to cloud-hosted models. While effective in isolation, these strategies remain either static, failing to adapt to the evolving inference behavior of external models, or costly, due to cryptographic or architectural constraints.  
144  
145  
146  
147  
148  
149  
150  
151  
152  
153  
154  
155  
156  
157  
158  
159  
160  
161  
162  
163  
164  
165  
166  
167  
168  
169

<sup>2</sup>An anonymized copy is attached with this submission for review.

## 2.2 Game-Theoretic Formulations

Game theory provides a principled foundation for analyzing strategic interactions in machine learning and multi-agent systems. Classical formulations such as zero-sum and Nash games underpin adversarial training and robust optimization, where competing agents iteratively minimize and maximize shared objectives (Goodfellow et al., 2014; Madry et al., 2017). Beyond these symmetric settings, hierarchical formulations, most notably Stackelberg games (Başar and Olsder, 1998; De Nittis and Trovo, 2016), model leader–follower dynamics and have become central to mechanism design, security, and privacy-aware learning. Recent studies have extended Stackelberg formulations to modern AI and large language models (LLMs). Theoretically, differentiable Stackelberg solvers now enable gradient-based optimization through implicit equilibria (Fiez et al., 2020; Li et al., 2020). In alignment, STA-RLHF (Makar-Limanov et al., 2024) interprets reinforcement learning from human feedback as a leader–follower interaction between the policy and reward model, while SGPO (Chu et al., 2025) leverages Stackelberg equilibrium for more data-efficient preference optimization. Beyond alignment, Stackelberg frameworks have been adopted for LLM detoxification (Xie et al., 2024), attacker–defender modeling in jailbreak prevention (Han and Zhu, 2025), and robust federated learning under adversarial or mixed attacks (Li et al., 2024).

## 3 Methodology

In this section, we present GTKA, a framework designed to balance quality and privacy in local-external LLM collaborative inference. We begin by formally defining the problem setting. Subsequently, we elaborate on the proposed method.

### 3.1 Problem Formulation

**Preliminaries.** We consider a setting involving two distinct agents: a *trusted local LLM* ( $\mathcal{M}_{loc}$ ) and an *untrusted external LLM* ( $\mathcal{M}_{ext}$ ).  $\mathcal{M}_{loc}$  operates within a secure boundary but is constrained by limited model parameters and knowledge. Conversely,  $\mathcal{M}_{ext}$  (e.g., GPT-5) possesses superior reasoning capabilities and dynamic knowledge but poses potential privacy risks as it resides in an external, untrusted environment.

**Task Definition.** Given a user’s sensitive query  $q$ , the primary objective is to acquire a high-quality an-

swer  $\hat{a}$  by leveraging  $\mathcal{M}_{ext}$ , without exposing the privacy of  $q$ . Directly querying  $\mathcal{M}_{ext}$  with  $q$  maximizes quality but compromises privacy, whereas relying solely on  $\mathcal{M}_{loc}$  ensures privacy but often yields suboptimal responses. Formally, we employ a trusted local LLM to generate a sequence of  $n$  low-leakage sub-queries  $\mathcal{S} = \{s_1, s_2, \dots, s_n\}$ . These sub-queries are sent to  $\mathcal{M}_{ext}$  to obtain a corresponding set of responses  $\mathcal{A} = \{a_1, a_2, \dots, a_n\}$ . Finally, a trusted local integrator aggregates the sub-queries, external responses, and original query  $(\mathcal{S}, \mathcal{A}, q)$  to synthesize the final answer  $\hat{a}$ .

### 3.2 The GTKA Framework

As illustrated in Figure 2, we propose a Game-theoretic Trustworthy Knowledge Acquisition (GTKA) framework. The framework consists of three modules: Privacy-Aware Sub-Query Generator, Adversarial Reconstruction Attacker, and Trusted Local Integrator.

#### 3.2.1 Privacy-Aware Sub-Query Generator

The privacy-aware sub-query generator aims to produce a set of low-leakage sub-queries that are maximally useful for downstream knowledge acquisition while minimizing exposure of the original sensitive intent. Let the privacy-aware sub-query generator be denoted as  $G_\theta$ , which is deployed within the trusted local environment.

Rather than paraphrasing the sensitive query  $q$ ,  $G_\theta$  generates a set of generalized, low-leakage sub-queries that preserve what information is needed to answer the request, such as key concepts to clarify, criteria to apply, constraints that do not reveal identities, and commonly used reasoning patterns, while masking sensitive entities and context-specific identifiers.

Specifically, for each user input  $q$ , the generator  $G_\theta$  samples  $K$  candidate sub-query groups  $\{\mathcal{S}^{(1)}, \dots, \mathcal{S}^{(K)}\}$ , where each group  $\mathcal{S}^{(k)} = \{s_1^{(k)}, \dots, s_n^{(k)}\}$  contains  $n$  sub-queries. For each candidate group  $\mathcal{S}^{(k)}$ , we send its sub-queries to the external LLM and obtain responses  $\mathcal{A}^{(k)} = \{a_1^{(k)}, \dots, a_n^{(k)}\}$ , where  $a_i^{(k)} = \mathcal{M}_{ext}(s_i^{(k)})$ . We then evaluate each group using a quality–privacy reward defined as

$$R(q, \mathcal{S}^{(k)}) = \alpha \text{Quality}(\hat{a}^{(k)}, a^*) - \beta \text{Leakage}(q, \mathcal{S}^{(k)}). \quad (1)$$

where  $\hat{a}^{(k)}$  denotes the locally integrated answer produced from  $(q, \mathcal{S}^{(k)}, \mathcal{A}^{(k)})$ ,  $a^*$  is the reference

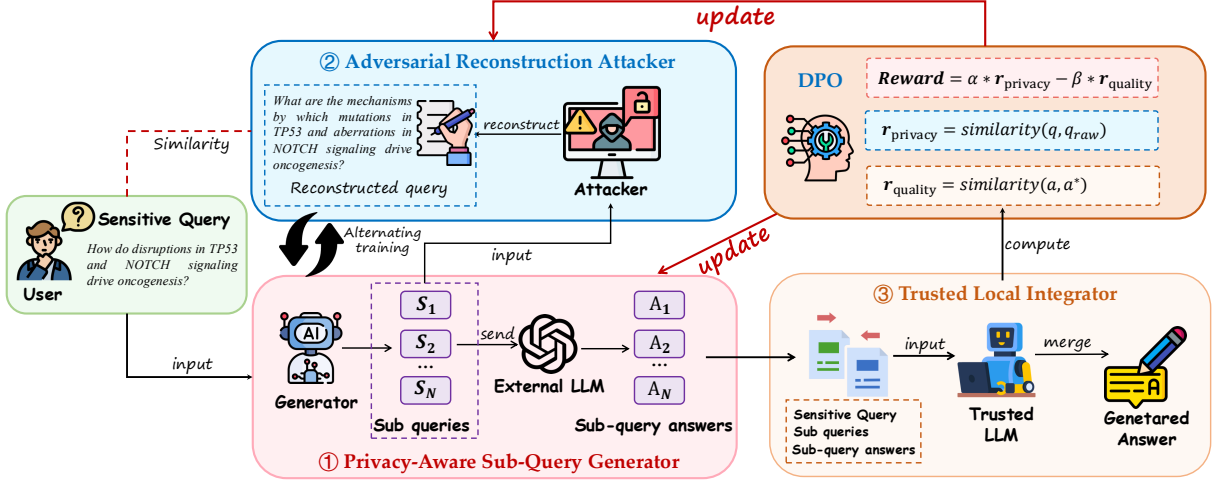


Figure 2: The overall framework of GTKA. The network consists of three modules: Privacy-Aware Sub-Query Generator, Adversarial Reconstruction Attacker, and Trusted Local Integrator.

answer, and  $\alpha, \beta$  control the quality–privacy trade-off.

The Quality term is defined as semantic similarity between the locally integrated answer  $\hat{a}^{(k)}$  and the reference answer  $a^*$ :

$$\text{Quality}(\hat{a}^{(k)}, a^*) = \text{Sim}(\hat{a}^{(k)}, a^*). \quad (2)$$

The leakage term  $\text{Leakage}(q, \mathcal{S}^{(k)})$  is provided by the adversarial reconstruction attacker described in Section 3.2.2. Based on this reward, we convert the  $K$  sampled candidate groups into pairwise preferences and train  $G_\theta$  with Direct Preference Optimization (DPO) (Rafailov et al., 2023).

**Preference Construction.** For each query  $q$ , we compute rewards  $\{R(q, \mathcal{S}^{(k)})\}_{k=1}^K$  for all sampled candidate groups and construct a preference pair by selecting a higher-reward group  $\mathcal{S}^+$  and a lower-reward group  $\mathcal{S}^-$ . Concretely,

$$\mathcal{S}^+ = \arg \max_k R(q, \mathcal{S}^{(k)}). \quad (3)$$

$$\mathcal{S}^- = \arg \min_k R(q, \mathcal{S}^{(k)}). \quad (4)$$

**DPO Optimization.** Let  $\pi_\theta(\mathcal{S} | q)$  denote the generator policy, and let  $\pi_{\text{ref}}(\mathcal{S} | q)$  be a fixed reference policy initialized from the pre-trained generator. Given each preference pair  $(q, \mathcal{S}^+, \mathcal{S}^-)$ , DPO updates  $G_\theta$  to assign higher likelihood to the preferred group  $\mathcal{S}^+$  than the less-preferred group  $\mathcal{S}^-$ .

### 3.2.2 Adversarial Reconstruction Attacker

To simulate realistic privacy threats, we introduce an Adversarial Reconstruction Attacker, denoted

as  $A_\phi$ . This module mimics an adversary who intercepts the sub-queries released to an untrusted external service and attempts to infer the user’s private intent. We formulate this attack as a reconstruction problem, where the attacker observes the set of sub-queries  $\mathcal{S}$  and predicts the most likely original query  $\hat{q}$ :

$$\hat{q} = A_\phi(\mathcal{S}) = \arg \max_{q'} P_\phi(q' | \mathcal{S}), \quad (5)$$

We quantify privacy leakage by measuring the semantic proximity between the reconstructed query  $\hat{q}$  and the ground truth  $q$ , where a higher similarity score indicates a more severe breach of privacy.

In the paper, the attacker  $A_\phi$  is implemented as an LLM and optimized via supervised fine-tuning. During GTKA training, the attacker provides an adaptive leakage signal that penalizes sub-queries that are easy to reconstruct. This adversarial mechanism encourages the generator to produce sub-queries that reveal less information and are more difficult to reverse-engineer into the original query. We train the generator and the attacker in an alternating manner. The detailed training procedure is presented in Algorithm 1.

### 3.2.3 Trusted Local Integrator

After the alternating training of the generator  $G_\theta$  and the attacker  $A_\phi$  converges, we fix the trained  $G_\theta$  and use it at inference time to generate a sub-query group  $\mathcal{S} = G_\theta(q)$  for each sensitive user query  $q$ . These sub-queries are then dispatched to the external LLM to obtain a set of responses  $\mathcal{A} = \{a_1, \dots, a_n\}$ , while the original query  $q$  remains strictly within the trusted local environment.

---

**Algorithm 1** GTKA Training

---

**Require:** Dataset  $\mathcal{D}$ , Generator  $G_\theta$ , Attacker  $A_\phi$ , Integrator  $\mathcal{I}$ , External LLM  $\mathcal{M}_{ext}$ .

**Ensure:** Optimized  $G_\theta$  and  $A_\phi$ .

- 1: Initialize  $\theta, \phi; \pi_{ref} \leftarrow G_\theta$ .
  - 2: **for** iteration  $t = 1, \dots, T$  **do**
  - 3:   Sample a batch of queries  $\mathcal{B}_q \subset \mathcal{D}$ .
  - 4:   // Step 1: Generate  $K$  candidates for each query in batch
  - 5:    $\mathcal{B}_S \leftarrow \{(q, \mathcal{S}^{(1..K)}) \mid q \in \mathcal{B}_q, \mathcal{S}^{(k)} \sim G_\theta(q)\}$ .
  - 6:   // Step 2: Optimize Attacker on generated samples
  - 7:   Update  $\phi \leftarrow \text{SFT}(A_\phi, \mathcal{B}_S)$  to maximize reconstruction likelihood.
  - 8:   // Step 3: Optimize Generator via DPO
  - 9:    $\mathcal{D}_{pref} \leftarrow \emptyset$
  - 10:   **for each**  $(q, \{\mathcal{S}^{(k)}\}_{k=1}^K) \in \mathcal{B}_S$  **do**
  - 11:     Compute rewards  $R^{(k)}$  for each candidate using **updated**  $A_\phi$ .
  - 12:     Construct pair  $(\mathcal{S}^+, \mathcal{S}^-)$ ; add to  $\mathcal{D}_{pref}$ .
  - 13:   **end for**
  - 14:   Update  $\theta \leftarrow \text{DPO}(G_\theta, \pi_{ref}, \mathcal{D}_{pref})$ .
  - 15: **end for**
  - 16: **return**  $G_\theta, A_\phi$
- 

After obtaining external responses, GTKA performs answer synthesis locally to ensure that the sensitive query  $q$  never leaves the trusted boundary. The trusted local integrator  $\mathcal{I}$  aggregates the original query, sub-queries, and external responses to produce the final answer:

$$\hat{a} = \mathcal{I}(q, \mathcal{S}, \mathcal{A}). \quad (6)$$

In our framework,  $\mathcal{I}$  is implemented as a trusted local LLM  $\mathcal{M}_{loc}$  that runs entirely within the secure boundary. It treats  $\mathcal{A}$  as external knowledge and synthesizes a final answer that directly addresses  $q$ . Only the sub-queries  $\mathcal{S}$  are sent to the untrusted external model, while the final reasoning that grounds external information in the user’s original intent is carried out locally. As a result, GTKA provides privacy by construction because  $q$  is never disclosed to  $\mathcal{M}_{ext}$ , yet it can still leverage  $\mathcal{M}_{ext}$  for broad and up-to-date knowledge.

## 4 Experiments

### 4.1 Dataset

To comprehensively evaluate the effectiveness and generalizability of our proposed method, we con-

duct experiments on two sensitive-domain QA datasets from the biomedical and legal domains, as both involve sensitive intents and high-stakes decision making where privacy leakage can cause real harm. Existing biomedical and legal QA datasets are often difficult to trace back to their source documents, which prevents rigorous evaluation of privacy leakage. Therefore, we construct two domain-specific QA datasets in this work, BioQA and LawQA.

BioQA is a high-quality question answering (QA) dataset in the biomedical domain, constructed from PubMed articles. For both datasets, we use an 8:2 split for training and testing. Dataset statistics are reported in Table 1, and the dataset construction process is detailed in Appendix A.

Type	BioQA	LawQA
Total	12,876	12,575
Train	10,301	10,060
Test	2,575	2,515
Avg. Q length (words)	13.90	128.18
Avg. A length (words)	52.42	146.95

Table 1: Statistics of the two datasets. “Avg. Q length” and “Avg. A length” denote the average number of words per question and answer, respectively.

### 4.2 Baselines and Evaluation Metrics

**Baselines. Local-Only (Trusted):** Answering the raw query solely with a local, trusted LLM (usually smaller) operating within the secure environment, without calling any external cloud LLMs. **External-Only (Untrusted):** Directly submitting the raw, unmodified sensitive query to the external cloud LLM. **PP-TS (Kan et al., 2023):** The framework protects user privacy by locally filtering sensitive information from user inputs before transmitting them to a remote LLM, and subsequently restoring the filtered content in the response. **IOI (Yao et al., 2024):** The true input instance is combined with a dummy instance to form an obfuscated instance, which is subsequently transmitted to the external LLM. **Hard-Prompt Obfuscation (Pape et al., 2025):** The method operates in the discrete token space, iteratively replacing tokens to generate a textually distinct yet functionally equivalent hard obfuscated prompt. **Soft-Prompt Obfuscation (Pape et al., 2025):** The method operates in the continuous embedding space, directly optimizing the prompt’s embedding

Method	Knowledge Acquisition (↑)					Intent Leakage(↓)		
	R-1	R-2	R-L	METEOR	BERTScore	ASR@1	ASR@3	MRR
Local-Only (Qwen2.5-3B-Instruct)	16.12	3.57	12.61	20.40	80.27	/	/	/
External-Only (Qwen-Turbo)	24.96	6.50	19.17	25.21	83.99	/	/	/
PP-TS (Kan et al., 2023)	16.58	3.63	12.78	20.56	80.52	21.48	39.65	31.60
IOI (Yao et al., 2024)	17.86	3.98	14.36	21.24	81.15	19.89	37.99	30.03
Hard-PO (Pape et al., 2025)	16.74	3.71	13.72	21.85	81.58	19.33	37.02	29.21
Soft-PO (Pape et al., 2025)	19.42	4.56	14.85	22.18	81.92	20.56	38.72	30.77
<b>GTKA</b>	<b>21.35</b>	<b>5.18</b>	<b>16.42</b>	<b>23.46</b>	<b>82.85</b>	<b>18.41</b>	<b>36.14</b>	<b>28.43</b>

Table 2: Performance comparison on the BioQA dataset with Qwen2.5-3B-Instruct (trusted local model) and Qwen-Turbo (untrusted external model). ↑ indicates higher scores are better, while ↓ indicates lower scores are better.

Method	Knowledge Acquisition (↑)					Intent Leakage(↓)		
	R-1	R-2	R-L	METEOR	BERTScore	ASR@1	ASR@3	MRR
Local-Only (Llama-3.1-8B)	16.35	3.26	11.62	18.17	81.02	/	/	/
External-Only (GPT-4o mini)	21.41	5.38	16.34	23.83	84.10	/	/	/
PP-TS (Kan et al., 2023)	15.62	3.25	11.76	19.24	81.67	14.16	31.98	25.76
IOI (Yao et al., 2024)	16.28	3.54	12.24	19.72	82.86	12.71	30.14	24.16
Hard-PO (Pape et al., 2025)	17.04	3.98	13.47	20.61	82.94	11.99	29.37	23.26
Soft-PO (Pape et al., 2025)	17.25	4.05	13.62	20.78	83.34	13.44	31.05	24.78
<b>GTKA</b>	<b>18.79</b>	<b>4.52</b>	<b>14.38</b>	<b>21.94</b>	<b>83.72</b>	<b>11.12</b>	<b>28.57</b>	<b>22.52</b>

Table 3: Performance comparison on the BioQA dataset with Llama-3.1-8B (trusted local model) and GPT-4o mini (untrusted external model). ↑ indicates higher scores are better, while ↓ indicates lower scores are better.

vectors to generate a functionally equivalent soft obfuscated prompt.

**Evaluation Metrics.** To evaluate our framework comprehensively, we assess all methods along two dimensions: knowledge acquisition and intent leakage, using both automatic and human evaluation. For Knowledge acquisition, the automatic evaluation reports commonly used QA metrics, including ROUGE-1 (R-1), ROUGE-2 (R-2), ROUGE-L (R-L), METEOR, and BERTScore.

For intent leakage, we adopt two attacker-based automatic metrics: Attack Success Rate (ASR) and Mean Reciprocal Rank (MRR). Specifically, for each privacy-preserved input, given a predefined candidate set consisting of the true source segment and several decoy segments, the attacker produces a ranked list of candidates according to how well each candidate matches the observed input. ASR@ $k$  is defined as Top- $k$  accuracy, i.e., the fraction of instances where the true segment appears in the top  $k$  positions. A higher ASR@ $k$  indicates stronger leakage. MRR measures the average reciprocal rank of the true segment in the attacker’s ranking. The detailed calculations are provided in Appendix B.

For human evaluation, we recruited three domain experts to annotate the data. For knowledge acquisition, the experts were provided with the question, the model-generated answer, and the gold reference answer, and then rated the response on a 0–5 scale along three dimensions: accuracy, completeness, and adoptability. For intent leakage, we conducted a human attack simulation: the experts were given the privacy-preserving query together with a candidate pool consisting of the true source segment and multiple decoy segments, and were asked to rank the candidate segments according to how likely each one was to be the true source based solely on the obfuscated query. We report the average score across the three experts as the final human evaluation result.

### 4.3 Experimental Settings

To comprehensively evaluate the performance of our proposed GTKA framework, we employ two distinct configurations of base Large Language Models (LLMs). First, we utilize Qwen2.5-3B-Instruct as the local trusted model paired with Qwen-Turbo as the external untrusted model, where the former serves as the backbone for both

Method	Knowledge Acquisition (↑)					Intent Leakage(↓)		
	R-1	R-2	R-L	METEOR	BERTScore	ASR@1	ASR@3	MRR
Local-Only (Qwen2.5-3B-Instruct)	39.99	17.90	28.74	43.94	81.57	/	/	/
External-Only (Qwen-Turbo)	45.84	21.06	32.49	45.33	85.80	/	/	/
PP-TS (Kan et al., 2023)	40.15	17.92	28.77	43.72	81.96	23.41	46.19	36.05
IOI (Yao et al., 2024)	40.72	18.08	29.02	43.83	82.41	21.94	44.34	34.41
Hard-PO (Pape et al., 2025)	41.24	18.25	29.18	44.03	82.35	21.24	43.49	33.70
Soft-PO (Pape et al., 2025)	42.56	18.42	29.35	44.56	83.87	22.57	45.22	35.07
<b>GTKA</b>	<b>43.52</b>	<b>19.86</b>	<b>30.84</b>	<b>44.91</b>	<b>84.69</b>	<b>20.35</b>	<b>42.56</b>	<b>32.74</b>

Table 4: Performance comparison on the LawQA dataset with **Qwen2.5-3B-Instruct** (trusted local model) and **Qwen-Turbo** (untrusted external model). ↑ indicates higher scores are better. ↓ indicates lower scores are better.

Method	Knowledge Acquisition (↑)					Intent Leakage(↓)		
	R-1	R-2	R-L	METEOR	BERTScore	ASR@1	ASR@3	MRR
Local-Only (Llama-3.1-8B)	40.49	18.64	28.78	43.97	81.81	/	/	/
External-Only (GPT-4o mini)	44.71	20.72	30.90	46.28	85.98	/	/	/
PP-TS (Kan et al., 2023)	38.56	16.82	26.54	42.28	82.13	25.43	48.46	37.92
IOI (Yao et al., 2024)	39.72	17.56	27.48	43.12	83.08	23.98	46.71	36.27
Hard-PO (Pape et al., 2025)	40.35	17.86	27.92	43.45	83.72	23.19	45.85	35.61
Soft-PO (Pape et al., 2025)	40.86	18.12	28.24	43.68	84.76	24.68	47.64	37.09
<b>GTKA</b>	<b>42.67</b>	<b>19.58</b>	<b>29.72</b>	<b>45.36</b>	<b>85.39</b>	<b>22.55</b>	<b>44.93</b>	<b>34.68</b>

Table 5: Performance comparison on the LawQA dataset with **Llama-3.1-8B** (trusted local model) and **GPT-4o mini** (untrusted external model). ↑ indicates higher scores are better, while ↓ indicates lower scores are better.

the generator and the attacker. Second, we adopt Llama-3.1-8B as the local trusted model and GPT-4o-mini as the external model, with Llama-3.1-8B similarly functioning as both the generator and attacker. Additional experimental settings are detailed in Appendix C.

#### 4.4 Experimental Results

Tables 2 and 3 report the results on the BioQA dataset under two different local-external model settings. Overall, GTKA maintains strong knowledge acquisition while achieving the lowest intent leakage. The performance on the LawQA dataset is detailed in Tables 4 and 5. The experimental results demonstrate that GTKA effectively masks sensitive legal intent while successfully obtaining pertinent content from external LLMs, thereby further validating the generalization capability of our framework in the legal domain.

To assess the practical usefulness and privacy protection of our method, we conducted a human evaluation, with results reported in Table 6. GTKA achieves the strongest performance across the accuracy, completeness, and adoptability metrics. For intent leakage, GTKA also exhibits substantially lower leakage than the compared method. These

findings indicate that the sub-queries generated by GTKA effectively mask the original intent from adversaries while still obtaining relevant knowledge from external LLMs and producing high-quality answers.

Human: Knowledge Acquisition (↑)			
Method	Accuracy	Completeness	Adoptability
Soft-PO	4.17	4.03	3.96
GTKA	4.65	4.54	4.72
Human: Intent Leakage (↓)			
Method	ASR@1	ASR@3	MRR
Soft-PO	24.25	42.84	33.65
GTKA	21.67	40.13	31.26

Table 6: Human evaluation results on the BioQA dataset. ↑ indicates higher scores are better, while ↓ indicates lower scores are better.

#### 4.5 Ablation study

To comprehensively evaluate the contribution of each component in our proposed framework, we conducted ablation studies on the BioQA dataset using Qwen2.5-3B-Instruct as the local model and Qwen-Turbo as the external model. The results are reported in Table 7, showing that each component contributes positively to the overall performance.

**Question:** How does canonical Wnt/ $\beta$ -catenin signaling regulate gene transcription in articular chondrocytes?

**Sub Query 1:**

What are the general mechanisms by which ligand-binding inhibits the proteasomal degradation of cytosolic proteins?

**Answer:**

When a ligand binds to its cell-surface receptor, it often triggers a signal transduction cascade that inhibits the cytosolic 'destruction complex'. Normally, this complex phosphorylates specific target proteins, marking them for ubiquitination and degradation by the proteasome.

Mechanism: Inhibition: **The mechanism proceeds as follows: 1. Inhibition: Ligand binding stops the phosphorylation activity. 2. Accumulation: The target protein escapes degradation, stabilizes, and accumulates in the cytoplasm (e.g., similar to how Wnt stabilizes  $\beta$ -catenin).**

**Sub Query 2:**

How do transcriptional co-activators that lack DNA-binding domains typically influence gene expression in the nucleus?

**Answer:**

These proteins act as bridges between DNA-bound factors and the transcription machinery. Since they cannot bind DNA directly, they operate through: 1. Protein-Protein Interaction: They bind to **sequence-specific transcription factors (TFs)** already on the promoter. 2. Chromatin Remodeling: They **recruit histone acetyltransferases like CBP/p300 to open up the chromatin structure.** 3. Repressor Displacement: They often **compete with and displace transcriptional co-repressors (e.g., Groucho/TLE)** to switch genes from 'off' to 'on'."

**Sub Query 3:**

What is the role of nuclear localization signals (NLS) in the translocation of stabilized cytosolic proteins?

**Answer:**

The Nuclear Localization Signal (NLS) is a specific amino acid sequence essential for transport. 1. Exposure: Once a cytosolic protein is stabilized, its NLS becomes accessible. 2. Transport: The NLS is recognized by **importins (karyopherins).** **These transport receptors bind the protein and ferry it through the Nuclear Pore Complex (NPC) into the nucleus,** allowing it to access genomic DNA.

**Generated Answer:**

By synthesizing the retrieved general mechanisms with the local context of articular chondrocytes, the pathway functions as follows: 1. Cytosolic Stabilization: **Wnt ligands initiate the process by inhibiting the proteasomal destruction complex. This prevents the ubiquitination of  $\beta$ -catenin, allowing it to accumulate.** 2. Nuclear Translocation: The stabilized  $\beta$ -catenin utilizes its NLS to interact with importins, traversing the Nuclear Pore Complex to **enter the nucleus.** 3. Transcriptional Regulation: Lacking its own DNA-binding domain,  $\beta$ -catenin **binds to TCF/LEF factors.** It activates gene expression by **facilitating the exchange of transcriptional repressors for histone acetyltransferases (HATs) including CBP/p300.** 4. Chondrocyte-Specific Outcome: This transcriptional machinery directly regulates the balance between anabolic matrix synthesis and catabolic enzyme activity (e.g., MMPs), thereby driving chondrocyte maturation and maintaining joint tissue homeostasis.

**Ground Truth:** When **Wnt ligands activate the pathway,  $\beta$ -catenin is stabilized** and **translocates to the nucleus,** where it **partners with TCF/LEF transcription factors.** This **complex displaces repressors, recruits coactivators such as CBP/p300,** and drives or represses target gene transcription that governs chondrocyte maturation, matrix synthesis, and catabolic enzyme production.

Figure 3: Biomedical case (Text highlighted in the same color denotes semantically corresponding information). Starting from the original query, our proposed GTKA method generates multiple low-leakage sub-queries that avoid revealing the original intent, submits them to an external LLM to obtain intermediate responses, and then uses a trusted local model to securely integrate these responses into the final answer.

Method	R-L	BertScore	ASR@3	MRR
GTKA (Full)	16.42	82.85	36.14	28.43
w/o DPO	15.73	82.19	38.45	30.12
w/o Quality ( $\alpha = 0$ )	14.47	81.13	33.24	25.69
w/o Privacy ( $\beta = 0$ )	17.08	83.07	40.32	33.41

Table 7: Ablation study on the BioQA dataset.

## 4.6 Case Study

We present a representative biomedical example to illustrate the advantage of our semi-permeable membrane framework, as shown in Figure 5. Given the original query, directly querying an external LLM would expose the pathway name, key molecule, and specific cell type. Instead, our method generates several low-leakage sub-queries that deliberately stay at a general-mechanism level while still covering the essential reasoning steps from different angles. The external model answers these generic sub-queries without being exposed to the original intent, and a trusted local model then securely integrates them back into the original biological context to produce the final explanation. The integrated answer aligns closely with the ground truth, demonstrating that our framework

can recover a highly consistent, context-specific response while minimizing intent leakage.

## 5 Conclusion

In this work, we presented Game-theoretic Trustworthy Knowledge Acquisition (GTKA) to reconcile the need for the advanced and continually updated knowledge of external LLMs with privacy constraints. By formulating query decomposition as an adversarial game, our framework generates generalized sub-queries that maximize utility while minimizing leakage. Through extensive experiments on two newly constructed benchmarks in the biomedical and legal domains, we demonstrated that GTKA significantly outperforms existing baselines, achieving a superior balance of privacy and answer fidelity. Unlike heavy cryptographic solutions or infrastructure-level changes, GTKA offers a lightweight, semantic-level defense that allows users to safely access the evolving information in external models without exposing proprietary intent, while remaining compatible with today's deployed LLM services and practical real-world workflows.

## 515 Limitations

516 Our study presents two primary limitations:

- 517 • While we have rigorously validated our  
518 method in the biomedical and legal domains,  
519 the framework’s generalizability to other  
520 fields remains to be verified. Sectors such  
521 as finance, engineering, or open-domain sce-  
522 narios possess distinct data distributions and  
523 privacy constraints. These differences could  
524 potentially shift the critical equilibrium be-  
525 tween utility and leakage.
- 526 • Our current approach dispatches sub-queries  
527 to a single external LLM provider. Theoretically,  
528 distributing sub-queries across multiple  
529 heterogeneous providers could offer stronger  
530 privacy guarantees. However, such a multi-  
531 provider strategy inevitably incurs higher com-  
532 putational overhead. We plan to explore opti-  
533 mization strategies that balance these resource  
534 demands with privacy gains in future work.

## 535 Ethical Statement

536 All experiments in this study were conducted in a  
537 controlled, simulated laboratory environment. We  
538 explicitly state that the datasets utilized in our work  
539 are sourced exclusively from publicly available  
540 benchmarks and contain no Personally Identifiable  
541 Information (PII). Furthermore, our adversarial as-  
542 sessments and privacy reconstruction attacks were  
543 strictly confined to local internal models. We did  
544 not perform any attacks against commercial exter-  
545 nal LLM servers or any third-party infrastructure.  
546 To facilitate reproducibility and transparency, all  
547 source code and data will be made publicly avail-  
548 able upon publication. For the review process, an  
549 anonymized version of the code and data has been  
550 submitted as supplementary material.

551 For the human evaluation conducted, we re-  
552 cruited three domain experts. These participants  
553 were compensated at a rate exceeding the local min-  
554 imum wage to ensure fair remuneration for their  
555 time and expertise. Prior to data collection, we  
556 explicitly explained the intended usage of the data  
557 to the annotators in detail. Regarding data consent,  
558 the BioQA and LawQA datasets constructed in  
559 this work utilize publicly available documents from  
560 PubMed and U.S. judicial decisions, respectively.  
561 These sources are open for research purposes; there-  
562 fore, individual consent was not required, and our

usage adheres to the data usage policies of the re-  
spective repositories.

## References

- Naafey Aamer, Muhammad Nabeel Asim, Shan Munir,  
and Andreas Dengel. 2025. Automating ai discovery  
for biomedicine through knowledge graphs and llm  
agents. *bioRxiv*, pages 2025–05.
- Martin Abadi, Andy Chu, Ian Goodfellow, H Bren-  
dan McMahan, Ilya Mironov, Kunal Talwar, and  
Li Zhang. 2016. Deep learning with differential pri-  
vacy. In *Proceedings of the 2016 ACM SIGSAC con-  
ference on computer and communications security*,  
pages 308–318.
- Tamer Başar and Geert Jan Olsder. 1998. *Dynamic  
noncooperative game theory*. SIAM.
- Amos Beimel. 2011. Secret-sharing schemes: A survey.  
In *International conference on coding and cryptology*,  
pages 11–46. Springer.
- Nicholas Carlini, Florian Tramer, Eric Wallace,  
Matthew Jagielski, Ariel Herbert-Voss, Katherine  
Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar  
Erlingsson, and 1 others. 2021. Extracting training  
data from large language models. In *30th USENIX  
Security Symposium (USENIX Security 21)*, pages  
2633–2650.
- Arup Kumar Chattopadhyay, Sanchita Saha, Amitava  
Nag, and Sukumar Nandi. 2024. Secret sharing: A  
comprehensive survey, taxonomy and applications.  
*Computer Science Review*, 51:100608.
- Kang Chen, Xiuze Zhou, Yuanguo Lin, Shibo Feng,  
Li Shen, and Pengcheng Wu. 2025. A survey on  
privacy risks and protection in large language mod-  
els. *Journal of King Saud University Computer and  
Information Sciences*, 37(7):163.
- Xu Chu, Zhixin Zhang, Tianyu Jia, and Yujie Jin.  
2025. Stackelberg game preference optimization for  
data-efficient alignment of language models. *arXiv  
preprint arXiv:2502.18099*.
- Badhan Chandra Das, M Hadi Amini, and Yanzhao Wu.  
2025. Security and privacy challenges of large lan-  
guage models: A survey. *ACM Computing Surveys*,  
57(6):1–39.
- Giuseppe De Nittis and Francesco Trovo. 2016. Ma-  
chine learning techniques for stackelberg security  
games: a survey. *arXiv preprint arXiv:1609.09341*.
- Tanner Fiez, Benjamin Chasnov, and Lillian Ratliff.  
2020. Implicit learning dynamics in stackelberg  
games: Equilibria characterization, convergence anal-  
ysis, and empirical study. In *International conference  
on machine learning*, pages 3133–3144. PMLR.

613	Craig Gentry. 2009. Fully homomorphic encryption using ideal lattices. In <i>Proceedings of the forty-first annual ACM symposium on Theory of computing</i> , pages 169–178.	
614		
615		
616		
617	Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. <i>arXiv preprint arXiv:1412.6572</i> .	
618		
619		
620	Zhengye Han and Quanyan Zhu. 2025. A dynamic stackelberg game framework for agentic ai defense against llm jailbreaking. <i>arXiv preprint arXiv:2507.08207</i> .	
621		
622		
623		
624	Priti Jagwani and Saroj Kaushik. 2017. Quality enhancement of location based services through real time context aware obfuscation using crowd sourcing. In <i>International Conference on Computational Science and Its Applications</i> , pages 79–94. Springer.	
625		
626		
627		
628		
629	Zhigang Kan, Linbo Qiao, Hao Yu, Liwen Peng, Yifu Gao, and Dongsheng Li. 2023. Protecting user privacy in remote conversational systems: A privacy-preserving framework based on text sanitization. <i>arXiv preprint arXiv:2306.08223</i> .	
630		
631		
632		
633		
634	Jiang Lai, Zheng Conghui, Zhang Xiaohan, Sun Fuhui, Wang Xiaoyan, and Pan Li. 2024. Leveraging llm based retrieval-augmented generation for legal knowledge graph completion. In <i>2024 IEEE 9th International Conference on Data Science in Cyberspace (DSC)</i> , pages 196–203. IEEE.	
635		
636		
637		
638		
639		
640	Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In <i>Advances in Neural Information Processing Systems</i> , volume 33, pages 9459–9474.	
641		
642		
643		
644		
645		
646		
647	Jiayang Li, Jing Yu, Yu Nie, and Zhaoran Wang. 2020. End-to-end learning and intervention in games. <i>Advances in Neural Information Processing Systems</i> , 33:16653–16665.	
648		
649		
650		
651	Tao Li, Henger Li, Yunian Pan, Tianyi Xu, Zizhan Zheng, and Quanyan Zhu. 2024. Meta stackelberg game: Robust federated learning against adaptive and mixed poisoning attacks. <i>arXiv preprint arXiv:2410.17431</i> .	
652		
653		
654		
655		
656	Lingjuan Lyu, Han Yu, and Qiang Yang. 2020. Threats to federated learning: A survey. <i>arXiv preprint arXiv:2003.02133</i> .	
657		
658		
659	Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. <i>arXiv preprint arXiv:1706.06083</i> .	
660		
661		
662		
663	Jacob Makar-Limanov, Arjun Prakash, Denizalp Goktas, Nora Ayanian, and Amy Greenwald. 2024. Sta-rlhf: Stackelberg aligned reinforcement learning with human feedback. In <i>Coordination and Cooperation for Multi-Agent Reinforcement Learning Methods Workshop</i> .	
664		
665		
666		
667		
668		
	Fan Mo, Hamed Haddadi, Kleomenis Katevas, Eduard Marin, Diego Perino, and Nicolas Kourtellis. 2021. Ppfl: Privacy-preserving federated learning with trusted execution environments. In <i>Proceedings of the 19th annual international conference on mobile systems, applications, and services</i> , pages 94–108.	669 670 671 672 673 674
	Olga Ohrimenko, Felix Schuster, Cédric Fournet, Aastha Mehta, Sebastian Nowozin, Kapil Vaswani, and Manuel Costa. 2016. Oblivious {Multi-Party} machine learning on trusted processors. In <i>25th USENIX Security Symposium (USENIX Security 16)</i> , pages 619–636.	675 676 677 678 679 680
	Sanjay Oruganti, Sergei Nirenburg, Jesse English, and Marjorie McShane. 2023. Automating knowledge acquisition for content-centric cognitive agents using llms. In <i>Proceedings of the AAAI Symposium Series</i> , volume 2, pages 379–385.	681 682 683 684 685
	Shuchao Pang, Zhigang Lu, Haichen Wang, Peng Fu, Yongbin Zhou, and Minhui Xue. 2024. Reconstruction of differentially private text sanitization via large language models. <i>arXiv preprint arXiv:2410.12443</i> .	686 687 688 689
	David Pape, Sina Mavali, Thorsten Eisenhofer, and Lea Schönherr. 2025. Prompt obfuscation for large language models. In <i>34th USENIX Security Symposium (USENIX Security 25)</i> , pages 2323–2342.	690 691 692 693
	Sanjaikanth E Vadakkethil Somanathan Pillai and Kiran Polimetla. 2024. Enhancing network privacy through secure multi-party computation in cloud environments. In <i>2024 International Conference on Integrated Circuits and Communication Systems (ICICACS)</i> , pages 1–6. IEEE.	694 695 696 697 698 699
	Maciej Pondeł, Iwona Chomiak-Orsa, Małgorzata Sobińska, Wojciech Grzelak, Artur Kotwica, Andrzej Małowiecki, Kamila Łuczak, Andrzej Greńczuk, Peter Busch, David Chudán, and 1 others. 2024. Ai tools for knowledge management—knowledge base creation via llm and rag for ai assistant. In <i>European Conference on Artificial Intelligence</i> , pages 3–15. Springer.	700 701 702 703 704 705 706 707
	Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. <i>Advances in neural information processing systems</i> , 36:53728–53741.	708 709 710 711 712
	Sofia Eleni Spatharioti, David Rothschild, Daniel G Goldstein, and Jake M Hofman. 2025. Effects of llm-based search on decision making: Speed, accuracy, and overreliance. In <i>Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems</i> , pages 1–15.	713 714 715 716 717 718
	Tan Su, Bingbing Zhang, Chi Zhang, and Lingbo Wei. 2024. Privacy leak detection in llm interactions with a user-centric approach. In <i>2024 IEEE 23rd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)</i> , pages 1647–1652. IEEE.	719 720 721 722 723 724

Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, 10(05):557–570.

Chong Wang, Mengyao Li, Junjun He, Zhongruo Wang, Erfan Darzi, Zan Chen, Jin Ye, Tianbin Li, Yanzhou Su, Jing Ke, and 1 others. 2025. A survey for large language models in biomedicine. *Artificial Intelligence in Medicine*, page 103268.

Xinhong Xie, Tao Li, and Quanyan Zhu. 2024. Learning from response not preference: A stackelberg approach for llm detoxification using non-parallel data. *arXiv preprint arXiv:2410.20298*.

Yixiang Yao, Fei Wang, Srivatsan Ravi, and Muhao Chen. 2024. Privacy-preserving language model inference with instance obfuscation. *arXiv preprint arXiv:2402.08227*.

Xuefei Yin, Yanming Zhu, and Jiankun Hu. 2021. A comprehensive survey of privacy-preserving federated learning: A taxonomy, review, and future directions. *ACM Computing Surveys (CSUR)*, 54(6):1–36.

Guoshenghui Zhao and Eric Song. 2024. Privacy-preserving large language models: Mechanisms, applications, and future directions. *arXiv preprint arXiv:2412.06113*.

## A Dataset Construction

For the BioQA dataset, we first collect 5,000 PubMed articles as source documents and use GPT-5 to generate three QA pairs per article, resulting in 15,000 raw QA pairs. To ensure data quality, we employ an LLM-based judge to score each QA pair on a 0–5 scale based on overall quality. We retain QA pairs with scores above 4 and further remove semantically redundant instances by filtering near-duplicate QA pairs. After these filtering steps, BioQA contains 12,876 QA pairs, with some illustrative examples shown in Table 8.

For the LawQA dataset, we collect 15,000 U.S. judicial decisions and use GPT-5 to generate one QA pair per decision, producing 15,000 raw QA pairs. We apply the same quality-scoring and near-duplicate removal procedure, retaining only QA pairs with scores above 4 and removing highly similar instances, yielding 12,575 QA pairs in the final LawQA dataset. Some illustrative examples are provided in Table 9.

## B Metric Calculations

We use ASR@k and MRR to quantify intent leakage under an attacker who observes the privacy-preserved input and attempts to recover the true

source segment from a candidate pool (the true segment plus decoys). ASR@k captures whether the attacker can successfully identify the true segment within the top- $k$  predictions, while MRR further reflects how highly the true segment is ranked across all candidates.

**Attack Success Rate (ASR).** For each test instance  $i \in \{1, \dots, M\}$ , let  $\mathcal{C}_i = \{c_{i,1}, \dots, c_{i,N}\}$  denote the candidate set containing exactly one true source segment  $c_i^*$  and  $N - 1$  decoy segments. Let the attacker output a ranked list over  $\mathcal{C}_i$ , and let  $r_i \in \{1, \dots, N\}$  denote the rank position of  $c_i^*$  (smaller is better). We define ASR@k as:

$$\text{ASR@k} = \frac{1}{M} \sum_{i=1}^M \mathbb{I}[r_i \leq k], \quad (7)$$

where  $\mathbb{I}[\cdot]$  is the indicator function.

Intuitively, ASR@k measures the probability that the attacker places the true segment among its top- $k$  candidates. A higher ASR@k indicates stronger intent leakage.

**Mean Reciprocal Rank (MRR).** MRR reflects how highly the true source segment is ranked within the attacker’s ranked list over the candidate set  $\mathcal{C}_i$ . A higher MRR indicates that the attacker tends to place the true segment closer to the top, suggesting stronger intent leakage, while a lower MRR indicates weaker leakage. Formally, MRR is defined as:

$$\text{MRR} = \frac{1}{M} \sum_{i=1}^M \frac{1}{r_i}, \quad (8)$$

where  $r_i$  denotes the rank position of the true segment  $c_i^*$  for instance  $i$ , and MRR takes values in  $(0, 1]$ .

## C Experimental Settings

Across all LLMs, we standardize the decoding hyperparameters with a sampling temperature of 0.7, nucleus sampling (Top- $p$ ) of 0.9, and a maximum generation length of 512 tokens. Regarding the specific hyperparameters of GTKA, we set the number of candidate sub-query groups  $K = 4$  and the number of sub-queries per group  $N = 9$ . The reward function balances quality and privacy with weights set to  $\alpha = 2/3$  and  $\beta = 1/3$ , respectively, and the iterative adversarial training process is conducted for 5 rounds. The generator is optimized via DPO

ID	Question	Answer
1	How do immune checkpoints like PD-1 and CTLA-4 suppress T-cell activation?	PD-1 and CTLA-4 attenuate T-cell receptor signaling by recruiting phosphatases that dephosphorylate key signaling molecules, reducing PI3K/Akt activity, IL-2 production, and co-stimulatory signals, which collectively limit T-cell proliferation and effector function.
2	How does NOD2 activation enhance dendritic cell maturation and antigen presentation in the context of vaccination?	NOD2 signaling upregulates costimulatory molecules (such as CD80 and CD86), MHC class II, and cytokines that promote T cell priming. It also induces autophagy, which facilitates antigen processing and can support cross-presentation to CD8+ T cells. These changes increase the efficiency of antigen presentation and promote migration to lymphoid tissues via chemokine receptor induction.

Table 8: Illustrative examples from BioQA.

ID	Question	Answer
1	In a Connecticut criminal matter where the court deferred the evidentiary portion of the defendant’s trial until the codefendant’s case concluded, I need to assess the viability of the defendant’s speedy-trial claim on appeal. Under Connecticut Practice Book § 3-10 and State v. Mooney, 218 Conn. 85 (1991), how should we frame the analysis given the record facts that the delay was caused by the prosecutor’s unforeseen back injury and that the defendant did not object to the deferral at trial? Please outline how these facts affect preservation and the merits, and what arguments the state can emphasize to show the delay was not sufficiently long and served the interest in a fair and impartial proceeding.	Applying Practice Book § 3-10 and Mooney’s balancing framework, the court weighed the defendant’s speedy-trial interest against the state’s interest in a fair, impartial trial and found no violation. Two facts are decisive. First, the delay was not significantly long, which undercuts the prejudice component of a speedy-trial claim. Second, the cause of delay—the prosecutor’s back injury—was unforeseen and beyond the state’s control, supporting the reasonableness of the deferral. On preservation, the defendant did not object to the delay at trial, so the speedy-trial claim was not preserved for appellate review. The state should emphasize that the deferral served the integrity of the proceedings, the length of delay was not substantial, the cause was legitimate and uncontrollable, and the absence of a contemporaneous objection bars the claim on appeal.
2	On the limitations issue, where the alleged offenses occurred in 1949 and the indictment was not found until 1952, and a motion to dismiss has been raised on statute-of-limitations grounds under 18 U.S.C. § 3282 (three-year limit), confirm whether the indictment is time-barred and indicate the proper disposition (including whether dismissal should be without prejudice) based on how the court applied the timing rule here.	Yes. Under 18 U.S.C. § 3282, an indictment must be found within three years after the commission of the offense. Because the offenses occurred in 1949 and the indictment was not found until 1952, it fell outside the three-year period and was time-barred. The court granted the motion to dismiss and dismissed the indictment without prejudice.

Table 9: Illustrative examples from LawQA.

with a KL penalty coefficient  $\beta$  of 0.1 and a learning rate of  $5e-6$ . The prompt templates utilized for generating sub-queries from the original query are detailed in Appendix D. For all comparative baselines, we strictly follow the optimal configurations reported in their original papers to ensure a fair comparison.

## D Prompt Details

The prompts used to generate sub-queries from the original queries on the BioQA and LawQA datasets are as follows:

**System Prompt:** You are a question generation expert specializing in privacy-preserving query generalization. Your goal is to generate  $n$  new, related, general-knowledge questions based on an original question, ensuring all specific entities and sensitive details are completely obfuscated. Follow these requirements:

1. **Abstraction Strategy:** Do not paraphrase or rewrite the original question. Instead, generate multiple new questions derived from broader concepts, underlying principles, mechanisms, and general knowledge in related domains.
2. **Strict Privacy Constraints:** The generated questions must not contain any specific identifiers. This strictly prohibits: lab/project names, specific gene/protein names, cell types, specific diseases, drugs, named biological processes, or concrete experimental data.
3. **Inference Utility:** Ensure the generated questions are valuable such that their combined answers allow a local model to infer the original answer, without exposing specific research targets to external models.

**Output Format:** Return only a valid JSON object.

**User Prompt:** Original question: {question}. Please generate  $n$  new general-knowledge questions.

Figure 4: Prompt (BioQA).

**System Prompt:** You are a legal question generation expert specializing in privacy-preserving query generalization. Your goal is to generate  $n$  new, related, general-knowledge questions based on an original legal question, ensuring all specific entities and sensitive details are completely obfuscated. Follow these requirements:

1. **Abstraction Strategy:** Do not paraphrase or rewrite the original question. Instead, generate multiple new questions derived from broader legal concepts, underlying principles, doctrines, and general standards in related domains.
2. **Strict Privacy Constraints:** The generated questions must not contain any specific identifiers. This strictly prohibits: party names, specific case numbers, docket numbers, judge names, geographic locations, or confidential case details.
3. **Inference Utility:** Ensure the generated questions are valuable such that their combined answers allow a local model to infer the original answer, without exposing specific case strategies or client matters to external models.

**Output Format:** Return strictly in JSON format only.

**User Prompt:** Original question: {question}. Please generate  $n$  new general-knowledge questions.

Figure 5: Prompt (LawQA).