# Query-Efficient Domain Knowledge Stealing Against Large Language Models

**Zhengao Li[1], Xiaopeng Yuan[2], Bolin Shen[1], Kien Le[1], Haohan Wang[3], Xugui Zhou[4], Shangqian Gao[1], Yushun Dong[1]**

[1]Florida State University,
[2]University of California, Los Angeles,
[3]University of Illinois Urbana-Champaign,
[4]Louisiana State University
zl23i@fsu.edu, xyuan75@ucla.edu, blshen@fsu.edu, kl23a@fsu.edu, haohanw@illinois.edu, xuguizhou@lsu.edu
sg24bi@fsu.edu, yushun.dong@fsu.edu

## Abstract

Large language models (LLMs) concentrate substantial knowledge in specialized domains due to extensive pretraining and instruction tuning, and they are now central to commercial and scientific practice. Yet access is usually limited to costly, rate-limited interfaces, which motivates methods that can extract targeted domain knowledge with minimal querying effort. A further challenge is that the target domain may be unknown in advance, so naive or generic prompts waste queries and fail to expose the underlying concepts and relations that structure the domain. In this work, we introduce a query-efficient approach for domain-specific knowledge stealing from black-box language models. Rather than issuing random questions or generic templates, our framework performs self-directed exploration that lets the model find the direction and mine domain knowledge by itself. Starting from a small and diverse seed, it discovers salient domain entities and induces their relations through structured question families that elicit definitional, functional, and compositional information. A feedback-driven controller analyzes the errors and uncertainty of the extracted surrogate model and uses this signal to refine subsequent queries, all without relying on prior domain knowledge or external resources. We evaluate the method in two expert-centric settings, medicine and finance, and observe consistently better performance while requiring significantly fewer queries.

## Introduction

Large language models (LLMs) have demonstrated substantial utility in high-stakes domains such as healthcare and finance. Domain-specialized LLMs, such as Med-PaLM 2 (Singhal et al. 2023) and BloombergGPT (Wu et al. 2023), deliver expert-level capabilities to clinicians and financial analysts. These models embody significant commercial value (Hosanagar and Krishnan 2024; Lin et al. 2024) and are increasingly deployed across a broad range of vertical applications (Wu et al. 2024; Jeong 2023). Consequently, they are often treated as critical intellectual property (IP) by their developers. To make these capabilities accessible, providers typically deploy such models via the LLM-as-a-Service (LLMaaS) paradigm (Yin et al. 2024; La Malfa et al. 2024), exposing them through cloud-hosted APIs (Zhang et al. 2024; Xu et al. 2025). However, this deployment model introduces significant security vulnerabilities. In particular, model stealing attacks have emerged as a pressing threat (Sha and Zhang 2024; He et al. 2024), in which adversaries systematically query public APIs to reconstruct a functionally equivalent surrogate model. This allows adversaries to reproduce the proprietary capabilities with minimal cost, effectively bypassing the intensive resources required for model development, and leading to unauthorized use of intellectual property that results in significant financial losses for the model owner.

Extensive research has revealed that LLMs are highly vulnerable to model stealing attacks (Aguilera-Martínez and Berzal 2025; Abdali et al. 2024), where adversaries can replicate a surrogate model simply by querying the target model (Sha and Zhang 2024). However, when targeting domain-specific LLMs, such attacks often require access to prior domain knowledge in order to craft effective queries and successfully replicate the model's specialized capabilities (Ran et al. 2025). Recent efforts in LLM stealing (Birch et al. 2023; Carlini et al. 2023; Finlayson, Ren, and Swayamdipta 2024) have achieved notable success, but they often depend on millions to tens of millions of API queries and rely heavily on domain-specific prompts as input. Approaches like Model Leeching (Birch et al. 2023) and EvoKD (Liu et al. 2024) have attempted to improve query efficiency via advanced prompt engineering or reinforcement learning. Nevertheless, these methods fundamentally assume access to the target domain and are incapable of stealing LLMs from completely unknown domains. However, in traditional machine learning, some recent works (Truong et al. 2021; Zhuang et al. 2024) have demonstrated the feasibility of model stealing without any domain knowledge by generating synthetic queries from noise distributions, successfully replicating proprietary models. However, directly transferring such strategies to LLMs is non-trivial due to the fundamental mismatch between continuous feature-based inputs in traditional models and the discrete nature of textual inputs in LLMs. As a result, stealing domain-specific LLMs without any manually provided domain knowledge remains an underexplored and technically challenging frontier.

However, stealing domain-specific knowledge from an

LLM remains exceptionally challenging due to several key obstacles. (1) *Lack of domain-specific entities*: Without any prior knowledge, it is difficult to even identify the core concepts, entities, or terminology that characterize the target domain. (2) *Difficulty in constructing relations*: Even if some domain-relevant entities are identified, constructing coherent and contextually grounded relationships among them remains a challenging task that is critical for formulating effective queries. (3) *Extensive Query Demand*: Effective model stealing typically requires a substantial number of queries to extract sufficient knowledge for training a high-fidelity surrogate model. The need for large-scale and informative querying becomes especially prohibitive when the generated queries are misaligned with the target domain. These challenges collectively highlight the inherent difficulty of stealing LLMs from entirely unfamiliar domains, where both query formulation and knowledge alignment remain largely underexplored and technically underdeveloped.

To address these challenges, we propose a novel framework that enables efficient domain knowledge stealing from LLMs without requiring manually provided domain knowledge or domain-specific prompts. Our method allows the model to autonomously explore and uncover domain-specific concepts, relationships, and terminology—thereby extracting specialized knowledge from the target model without external guidance. Specifically, our method consists of three key components: First, we employ an active query exploration mechanism to rapidly probe the target model and discover domain-relevant entities. We initialize this process with a small set of domain-agnostic prompts that do not contain domain-specific terms. This step enables the generation of highly informative and domain-aligned queries, even without explicit domain-specific supervision. Second, we extract answers from the target model and segment them into fine-grained textual chunks, from which we extract lightweight semantic cues (e.g., SBERT similarity) to group related entities. This chunk-level analysis helps progressively approximate the structure of the target domain. Finally, we introduce a Chain-of-Question mechanism that adaptively composes follow-up queries to improve coverage and stability. By leveraging the surrogate model's current understanding, we dynamically refine the querying process to actively acquire uncertain or unexplored knowledge regions. Together, these components form a largely automated domain knowledge stealing pipeline that helps reduce the overall query cost compared with prior methods while enabling faithful imitation of domain-specific behaviors.

Our contributions can be summarized in three-fold:

- **Domain Knowledge Stealing.** To the best of our knowledge, we are among the first to study the challenge of stealing domain-specific knowledge from a target LLM without relying on manually provided domain knowledge or domain-specific prompts.
- **Self-Directed Knowledge Discovery.** We propose a framework that enables the model to autonomously uncover domain knowledge, including identifying domain-specific entities and extraction lightweight semantic cues that link related concepts without external guidance.
- **Comprehensive Evaluation.** Extensive experiments across Medicine and Finance show that our method achieves strong performance and competitive query efficiency compared with prior approaches.

# Preliminaries

## Notations

We use lowercase letters (e.g., $a$) to denote scalar variables, and bold lowercase letters (e.g., $\mathbf{b}$) to represent vectors. Calligraphic uppercase letters (e.g., $\mathcal{Q}$) denote sets, and we use $q$ to represent a single query instance. A query $q$ is instantiated using a domain entity $e$ and a semantic prompt type $t$, such that $q = G(e, t)$, where $G(\cdot)$ is a prompt generation function. The target and surrogate models are denoted by $T$ and $S$, respectively.

## Problem Statement

**Model Stealing.** We study model stealing against large language models (LLMs) under black-box API access (Orekondy, Schiele, and Fritz 2019; Sun et al. 2024). The adversary can query a proprietary target model $T$ and aims to construct a surrogate model $S$ that imitates its behavior. The adversary has no access to the internal architecture, training data, or manually provided domain-specific information used by $T$, and can only observe the output responses returned by the API. Formally, given an input query $q \in \mathcal{Q}$, the adversary receives $T(q)$ and accumulates a query–response dataset $\{(q, T(q))\}_{q \in \mathcal{Q}}$ to train $S$. The goal is for $S$ to approximate the behavior of $T$ over unseen queries.

**Domain Knowledge.** LLMs deployed in specialized fields (e.g., medicine, finance) implicitly encode extensive domain-specific knowledge, including entities, hierarchical concepts, and associations among them. In this paper, we consider a challenging setting in which the adversary has no manually provided domain knowledge—such as predefined entity lists, knowledge bases, or exemplar documents—to guide the stealing process. Instead, the adversary must rely solely on black-box querying to uncover domain-relevant entities and extract lightweight semantic cues that relate them. Our objective is to develop a black-box stealing framework that can progressively approximate the domain-relevant knowledge expressed by the target model through its responses.

# Methodology

Our objective is to extract domain-specific knowledge from a black-box large language model (LLM) under strict query constraints, without relying on labeled data or predefined ontologies. To this end, we propose a fully automated, data-free distillation framework that incrementally constructs a high-coverage question-answer corpus tailored to the target domain. The framework tackles three fundamental challenges: (1) identifying domain-relevant entities in the absence of supervision, (2) uncovering implicit semantic relationships among those entities, and (3) minimizing query overhead while enabling the training of a high-fidelity surrogate

model. Figure 1 illustrates the overall pipeline, which comprises three key stages: Identifying Domain-Specific Entities, Discovering Implicit Semantic Relations, and Query-Efficient Exploration via Perplexity-Guided Scheduling.

## Identifying Domain-Specific Entities

The first challenge arises from the lack of prior domain knowledge: without any labeled data or external resources, it is unclear what entities, terminology, or concepts are relevant to the target domain. To overcome this, we propose a fully automated pipeline that discovers domain-specific entities through interaction with the black-box model itself. This is achieved through iterative prompting, entity extraction, and lightweight relevance filtering applied directly to the target model's responses. We begin by issuing a small set of open-ended, domain-agnostic prompts to $T$, which encourage the model to expose the terminology it considers central to the underlying domain. From these responses, we extract candidate domain entities using a pretrained named-entity extractor. We filter responses that exhibit low sentence-level semantic coherence using a coherence-based heuristic `answer_is_broad`, which detects responses whose consecutive sentences are semantically distant or mix unrelated topics. Formally, we define the initial entity set as

$$\mathcal{E}_d^{(0)} = \bigcup_{i=1}^{n} \text{ExtractEntities}(T(\texttt{Prompt}_i)).$$

For each discovered entity $e \in \mathcal{E}_d^{(0)}$, we generate structured queries using four canonical prompt templates corresponding to semantic roles: DEF (definition), CAT (classification), FUN (function), and PART (part-of). This equation yields a seed query pool:

$$\mathcal{Q}_0 = \left\{ G(e,t) \mid e \in \mathcal{E}_d^{(0)},\ t \in \{\text{DEF}, \text{CAT}, \text{FUN}, \text{PART}\} \right\}.$$

The choice of these four templates is grounded in converging evidence from ontology engineering, lexical semantics, question classification, and educational theory. Ontology frameworks commonly rely on primitive relations such as IS_A, PART_OF, and HAS_FUNCTION (Smith, Arp, and Spear 2015), which map directly to our prompts for definition, composition, and function. Lexical semantics, particularly in the Generative Lexicon framework (Pustejovsky 1995), identifies the *Formal*, *Constitutive*, and *Telic* qualia roles as essential to concept representation, aligning closely with our semantic templates. Additionally, standard QA taxonomies treat definition, list, purpose, and meronymy as distinct and sufficient classes for factual inquiry (Li and Roth 2002; Craswell et al. 2020). Finally, pedagogical models like Bloom's taxonomy and SOLO taxonomy recognize these four dimensions as the minimal set required for conceptual mastery (Bloom 1956; LW et al. 2001; Biggs and Collis 1982). Taken together, these perspectives validate our use of these four prompt types as both semantically complete and practically efficient for exploring domain knowledge. Although $\mathcal{E}_d^{(0)}$ is constructed without domain supervision, we apply a simple but effective constraint to maintain topical relevance. Specifically, we issue simple refinement prompts (e.g., "list only the domain-specific concepts") and retain only entities that consistently reappear in these constrained responses, ensuring topical relevance without external supervision. Only entities confirmed by these constrained responses are retained. This iterative refinement yields a domain-aligned entity set $\mathcal{E}_d$ that reflects the terminology emphasized by $T$, without relying on external corpora or ontologies. It serves as the foundation for subsequent relation discovery and knowledge extraction.

Algorithm 1: Query-Efficient Stealing Loop

**Require:** Initial seed entities $\mathcal{E}_d$ from target domain $d$, surrogate model $S$, target model $T$, query budget $B$, correction threshold $\tau_{\text{corr}}$, clustering interval $R$

Initialize query pool $\mathcal{Q} \leftarrow$ Generate four prompts (DEF, CAT, FUN, PART) for each $e \in \mathcal{E}_d$
**for** $t = 1$ to $B$ **do**
  $q^\star \leftarrow \arg\max_{q \in \mathcal{Q}} \text{PPL}_S(q)$ // Select the question with highest perplexity to target the weakest area
  $(a_T, a_S) \leftarrow (T(q^\star), S(q^\star))$ // Query both target and surrogate models to compare responses
  Add $(q^\star, a_T)$ to training dataset
  **if** SBERT_cos$(a_T, a_S) < \tau_{\text{corr}}$ **then**
    Generate correction prompt using $a_S$ and $q^\star$, then query $T$ for explanation
    Add explanation and corrected answer to training cache
  **end if**
  **if** answer_is_broad$(a_T)$ **then**
    Segment $a_T$ into coherent chunks via SBERT and clustering
    Generate one follow-up question per chunk and enqueue them into $\mathcal{Q}$
  **end if**
  **if** $t \bmod R = 0$ **then**
    Encode recent $a_T$ answers via SBERT and cluster them
    Identify least-covered cluster and enqueue a representative query from it
  **end if**
  Update $S$ using current training dataset
**end for**
**return** Trained surrogate model $S$

## Discovering Implicit Semantic Relations

The second challenge concerns the difficulty of constructing meaningful relationships between discovered entities. While isolated facts can be extracted through static prompting, a coherent representation requires capturing distinctions or contextual conditions that connect related concepts. To this end, we develop a two-pronged strategy that leverages the black-box model's own feedback to uncover and refine latent semantic structure. We detect discrepancies between the surrogate and target answers using SBERT-based semantic similarity. Let $q^\star$ denote the current highest-priority question according to perplexity-based sampling. After obtain-

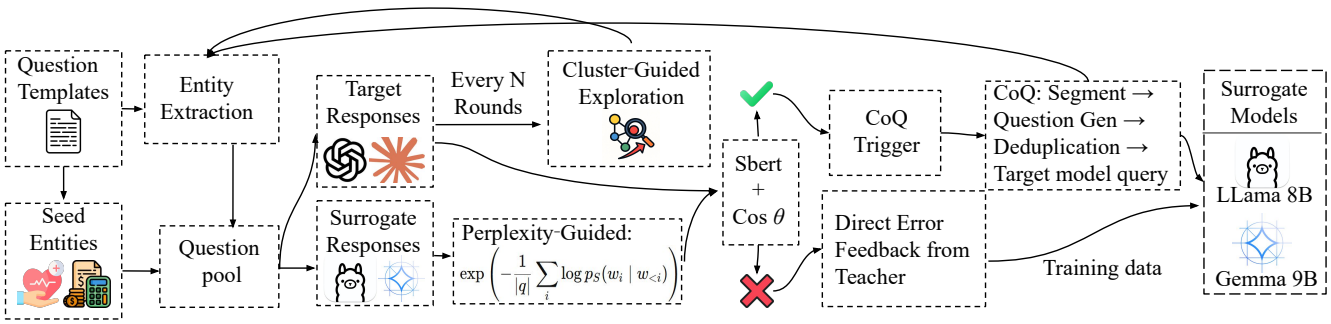



Figure 1: Overview of our query-efficient domain knowledge extraction pipeline. We begin with structured question templates and extract seed entities from the target model's responses. These are used to generate an evolving question pool, which is queried against the target LLM to obtain answers. The surrogate model answers the same questions and we rank them by the perplexity of its generated outputs. Questions with high perplexity or large SBERT-based semantic deviation are sent to the target for correction or CoQ-style follow-up. Corrected answers are added to the surrogate's training cache. Cluster-guided exploration is triggered every $N$ rounds to improve conceptual coverage.

ing the answer pair $(a_T, a_S) = \big(T(q^\star), S(q^\star)\big)$, we measure semantic divergence via SBERT-based cosine similarity (Reimers and Gurevych 2019). When the similarity falls below a pre-defined threshold $\tau_{\text{corr}}$, we prompt the target model to explain the error, producing a targeted corrective prompt such as: `Explain why` $S(q^\star)$ `is incorrect and what the correct answer is.` The target's explanation typically highlights details the surrogate missed, such as subtype differences, causal factors, or contextual conditions. From these explanations, we extract new queries using the same entity extraction and templating logic as in Section . This corrective feedback loop not only improves surrogate coverage but also reveals subtle distinctions between closely related concepts. Complementary to this is our chunk-based CoQ (Chain-of-Questioning) mechanism, which transforms vague or overly broad answers into finer-grained knowledge. To detect such cases, we define a heuristic answer_is_broad$(a) = \mathbb{1}[\min_i \cos(s_i, s_{i+1}) < \tau]$, where $s_i$ and $s_{i+1}$ are consecutive sentence embeddings from a pre-trained SBERT model, and $\tau$ is a coherence threshold. If any two adjacent sentences are semantically disjoint, such answers are treated as overly broad because they bundle multiple topics together, which triggers chunk-based follow-up. When $T(q^\star)$ is flagged as overgeneral, we segment the response into logical chunks and construct a follow-up question for each. These follow-up questions typically take the form of elaborations on specific answer segments (e.g., `Can you elaborate on: ''chunk_i''?`).

To help visualize our pipeline, we consolidate key illustrative examples across four modules into a unified walkthrough. This includes structured prompting, correction, fine-grained follow-up, and entity extraction. This recursive refinement breaks high-level or multi-topic answers into more specific components, enabling finer-grained coverage of the domain, especially when the model's initial answer spans multiple distinct concepts. Together, the correction-based prompting and chain-of-questioning routines constitute a model-guided mechanism for uncovering lightweight semantic cues that indicate how concepts relate or differ. Rather than relying on a predefined schema, we rely on the model's own explanations and decompositions to reveal useful distinctions, which in turn enables adaptive and grounded discovery of domain relationships.

**Illustrative Prompt Examples**

**Seed Entity: Anaphylaxis** *Definition Prompt:* What is anaphylaxis in clinical medicine?
*Classification Prompt:* What are the major types or categories of anaphylaxis?
*Function Prompt:* What is the physiological role or consequence of anaphylaxis?
*Part-of Prompt:* Anaphylaxis is part of which broader medical condition or syndrome?
**Model Feedback-Driven Prompts:**
*Correction Prompt:* The surrogate answer incorrectly states that anaphylaxis is a chronic condition. Can you explain the correct nature of anaphylaxis?
*Chunked CoQ Prompts:*
- What are the typical symptoms of anaphylaxis during the early phase?
- How does anaphylaxis affect the respiratory system?
- What is the standard emergency treatment for severe anaphylaxis?

## Query-Efficient Exploration via Perplexity-Guided Scheduling

The final challenge arises from the high query cost in black-box extraction. Since each API call incurs cost and latency, it is essential to allocate a limited query budget toward interactions that are expected to provide the largest improvement to the surrogate. We therefore design a scheduling strategy centered around the perplexity of the surrogate model's *generated answer*, which we use as a lightweight signal of uncertainty. At iteration $t$, let $a_S(q)$ denote the surrogate-generated answer to a question $q$. We compute its answer perplexity $\text{PPL}_S(a_S(q))$ and prioritize questions with higher perplexity values. Formally, the next query is

selected as

$$q^\star = \arg\max_{q \in \mathcal{Q}} \mathrm{PPL}_S(a_S(q)).$$

Although perplexity is not a direct measure of epistemic uncertainty, higher answer perplexity often coincides with cases where the surrogate lacks stable lexical or semantic expectations. We therefore treat it as a simple, model-agnostic proxy for identifying queries that the surrogate finds most difficult. This ensures that the surrogate is consistently challenged with inputs that expose gaps in its current knowledge distribution, maximizing learning gain per query. To further encourage conceptual breadth, we introduce a periodic diversity booster. Every $R$ iterations, we collect a fixed-size window of recent surrogate outputs, embed them using SBERT, and apply a standard distance-based clustering method (e.g., $k$-means). We identify the smallest cluster—corresponding to concepts that have received the least attention—and promote one of its questions to the top of the queue. Formally, if $\mathcal{C} = \{C_1, \ldots, C_k\}$ denotes the clusters, we sample a representative $q$ from $\arg\min_i |C_i|$ and schedule it next.

This dual mechanism—combining uncertainty-driven prioritization with diversity-aware reranking—ensures balanced depth and coverage in knowledge extraction. Rather than passively sampling or following hand-designed trajectories, the surrogate actively guides interactions toward regions where its representation is weakest. As shown in Algorithm 1, this scheduling strategy integrates with entity discovery and relation refinement, enabling higher-fidelity surrogates under tight query budgets.

## Experimental Evaluations

To guide our empirical study, we organize our evaluation around the following research questions (RQs): **RQ1 (Effectiveness):** How well does our surrogate model perform on domain-specific QA tasks compared to strong black-box baselines? **RQ2 (Query Efficiency):** Does our method match or exceed prior approaches while using fewer queries? **RQ3 (Ablation Study):** How much does each component contribute to the overall performance?

**Downstream Task and Datasets.** To evaluate whether our query-efficient framework improves downstream performance under constrained budgets (RQ1), we conduct experiments on six question-answering benchmarks across two knowledge-intensive domains: medicine and finance. In the medical domain, we use MedQA (Jin et al. 2020), a benchmark of professional clinical multiple-choice exam questions; PubMedQA (Jin et al. 2019), a biomedical factual QA dataset based on research abstracts; and ChemProt (Kringelum et al. 2016), a chemical–protein relation classification task. In the financial domain, we use FOMC (Shah, Paturi, and Chava 2023), which evaluates understanding of U.S. Federal Reserve monetary policy texts; HeadLine (Sinha and Khandait 2020), a causal inference dataset for financial headlines; and FPB (Malo et al. 2014), a phrase-level sentiment classification task from financial analyst reports. We compare the following systems: (1) two untuned base models: LLaMA-3.1 8B Instruct and Gemma-9B Instruct; (2) domain-specific targets, obtained by fine-tuning LLaMA-3.1 8B on domain-relevant corpora; (3) two black-box baselines, EvoKD (Liu et al. 2024) and Model Leeching, both constrained to the same query budget; and (4) our LoRA surrogate, trained using QA pairs produced by our structured, query-efficient distillation framework. All results are reported using accuracy, measuring the percentage of correctly answered instances. This unified metric allows fair comparison across datasets and systems, and directly reflects the effectiveness of each method in improving downstream task performance (Hu et al. 2021). This evaluation setup addresses RQ1 by testing whether our method delivers stronger task-specific performance under the same black-box querying constraints.

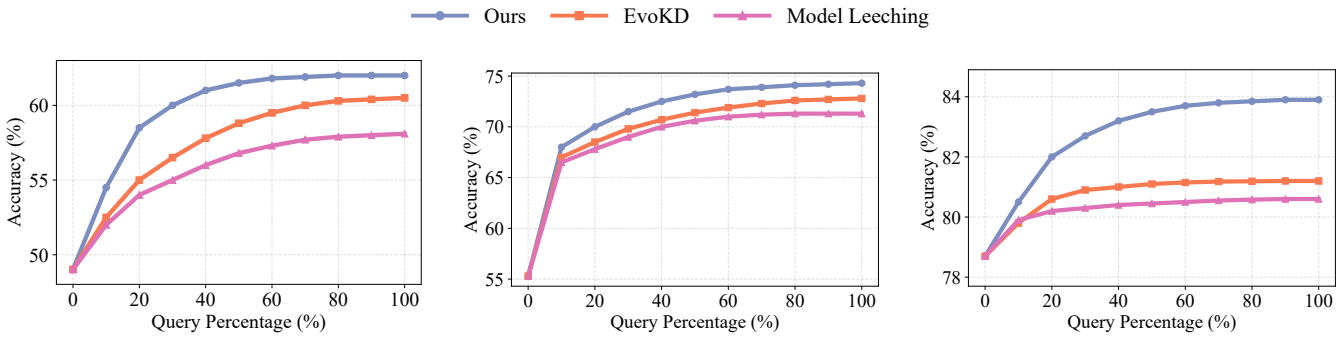
**Baselines.** To ensure a fair and controlled evaluation, we re-implement two representative black-box distillation baselines: *Model Leeching* (Birch et al. 2023) and *EvoKD* (Liu et al. 2024). Our goal is to isolate the effects of prompt design and query scheduling under identical model, budget, and optimization conditions. For **Model Leeching**, we follow Birch et al., where the attacker passively queries the target LLM with natural-language prompts and collects the answers without feedback or refinement. To reproduce this setup, we generate 500 simple QA-style prompts using only the question texts from two domain-relevant public datasets: MedMCQA (medical QA) and finance-related subsets of MMLU (e.g., macroeconomics, microeconomics, accounting). These datasets are used solely as sources of question templates; their labels are never used. This preserves the passive nature of Model Leeching and ensures that any performance differences arise from querying strategy rather than domain supervision. For **EvoKD**, we implement the entropy-guided scheduling strategy proposed by Liu et al. To make the comparison focus solely on query-ordering effects, we keep EvoKD's iterative selection and conversational structure, but replace its free-form prompt generation with the same QA-style seed prompts used above. Thus, both methods receive identical prompt content, and only the prioritization mechanism differs. Across both baselines, we fix the surrogate architecture, optimization hyperparameters, 500-query budget, and evaluation protocol to match our framework. This consistent setup ensures that observed differences in performance directly reflect the querying strategy and structured prompt design.

### Effectiveness of Domain Knowledge Stealing

To evaluate **RQ1**, we examine whether our surrogate models, trained with only 500 high-quality QA pairs, can effectively extract domain-specific knowledge from black-box LLMs and outperform existing baselines across six QA benchmarks. As shown in Table 1, our method consistently outperforms both baselines under the same query budget, indicating that uncertainty-guided scheduling uses each query more efficiently during surrogate training. In the medical domain, our LLaMA-3.1-based surrogate reaches 62.0% on MedQA, 74.9% on PubMedQA, and 43.1% on ChemProt, outperforming EvoKD on all three tasks (60.5%, 72.8%, 40.3%) while remaining close to Model Leeching on ChemProt. Our Gemma-2-based surrogate further advances performance, achieving 64.6% on MedQA, 67.3%

| Method / Model | Medicine | | | Finance | | |
|---|---|---|---|---|---|---|
| | MedQA | PubMedQA | ChemProt | FOMC | HeadLine | FPB |
| ChatGPT-4.1 (Target Model) | 75.5 | 80.1 | 64.9 | 67.1 | 88.3 | 73.0 |
| LLaMA-3.1 8B (Base) | 49.0 | 55.3 | 24.7 | 41.0 | 78.7 | 65.9 |
| EvoKD | 60.5±0.7 | 72.8±0.9 | 40.3±1.0 | 43.5±1.0 | 80.9±1.1 | 66.0±1.0 |
| Model Leeching | 58.1±0.2 | 74.3±0.9 | **44.7±0.4** | 48.5±1.1 | 80.4±1.0 | 66.7±0.5 |
| **Ours** | **62.0±1.0** | **74.9±0.8** | 43.1±0.4 | **54.0±1.0** | **83.9±0.9** | **68.8±0.9** |
| Gemma-2 9B (Base) | 57.1 | 54.6 | 20.8 | 49.8 | 77.4 | 41.9 |
| EvoKD | 62.0±0.3 | 61.8±0.4 | 39.9±0.3 | 49.5±0.6 | 78.5±0.4 | 57.9±0.5 |
| Model Leeching | 60.2±0.4 | 63.6±0.5 | 42.4±0.4 | 50.1±0.5 | 79.0±0.6 | 60.1±0.4 |
| **Ours** | **64.6±0.2** | **67.3±0.5** | **44.5±0.3** | **54.2±0.4** | **80.4±0.2** | **68.5±0.3** |

Table 1: Accuracy (%) of all methods on six QA benchmarks across Medicine and Finance.




(a) Medical: MedQA (b) Medical: PubMedQA (c) Finance: HeadLine

Figure 2: Query efficiency comparison across different domains. Our method consistently achieves faster convergence and higher final accuracy.

on PubMedQA, and 44.5% on ChemProt, clearly outperforming the base model by large margins and surpassing both baselines on all three tasks. In the financial domain, our LLaMA-based surrogate again delivers the highest accuracy across all three financial benchmarks, obtaining 54.0% on FOMC, 83.9% on HeadLine, and 68.8% on FPB, improving over EvoKD (43.5%, 80.9%, 66.0%) and Model Leeching (48.5%, 80.4%, 66.7%). Similarly, our Gemma-based surrogate outperforms its base by wide margins and exceeds both distillation baselines, reaching 54.2% on FOMC, 80.4% on HeadLine, and 68.5% on FPB. These results indicate that our method enables effective domain knowledge transfer even under strict black-box conditions, consistently surpassing both untuned foundations and strong distillation alternatives across diverse domains.

## Efficiency of Domain Knowledge Stealing

To evaluate **RQ2**, we examine whether our method can reach high accuracy with fewer target queries, reflecting stronger query efficiency compared with prior approaches. As shown in Figure 2, we measure performance on three representative datasets—MedQA, PubMedQA, and HeadLine—by gradually increasing the number of QA pairs from 0% to 100% of the 500-query budget. For fairness, all methods are trained with the same LoRA surrogate, the same optimization setup, and the same 500-query constraint. Under each subset size, we fine-tune surrogate models using our method, EvoKD, and Model Leeching under matched configurations. Across all stages of the query process, our approach consistently achieves higher accuracy and faster convergence than both baselines. On MedQA, our model reaches 59.5% accuracy with only 30% of the queries, already surpassing EvoKD and Model Leeching at their full budgets, and continues to climb to over 62% as more queries are used. On PubMedQA, the performance gap emerges even earlier, with our method achieving over 74% accuracy at 60% of the budget while both baselines plateau below 73%. A similar trend is observed on HeadLine, where our approach quickly reaches 84% accuracy with just 50% of the queries, while EvoKD and Model Leeching converge more slowly and saturate around 81%. These results demonstrate that our framework not only achieves better final accuracy, but also converges significantly faster, offering a more efficient use of queries under black-box constraints.

## Ablation Study

To address **RQ3**, we conduct ablation experiments on two representative datasets—MedQA and FOMC—to quantify

| Task | Full | w/o Cluster | w/o CoQ | w/o FB |
|---|---|---|---|---|
| MedQA | 62.0 | 60.8 | 59.1 | 58.9 |
| FOMC | 54.0 | 53.4 | 52.2 | 53.9 |

Table 2: Ablation study on MedQA and FOMC.

the individual contributions of core components in our framework. As shown in Table 2, removing any one of the three modules leads to performance degradation, confirming their necessity. The most substantial drop occurs when CoQ is removed, with MedQA falling from 62.0% to 59.1% and FOMC from 54.0% to 52.2%, highlighting the critical role of structured multi-turn reasoning in extracting fine-grained domain knowledge. Disabling Feedback-Guided Scheduling (FB) also harms performance, particularly on FOMC, where accuracy decreases to 53.9%, suggesting that uncertainty-aware prompt prioritization significantly improves sample efficiency in decision-heavy tasks. Removing Cluster-Guided Exploration results in a more moderate decline, indicating that while topic clustering enhances coverage diversity, it plays a more supportive role compared to the other modules. These findings imply that components promoting follow-up reasoning and adaptive sampling contribute more directly to knowledge acquisition than purely coverage-based strategies, especially in complex domains like medicine and finance where useful knowledge often lies in nuanced or sparsely distributed contexts.

## Related Work

Research on black-box model extraction has expanded rapidly alongside the development of large language models (LLMs). In black-box settings, an adversary interacts with a proprietary model only through its outputs, and attempts to reconstruct a surrogate that captures the target's behavior. Prior work spans multiple directions, including function cloning, knowledge distillation, and active query design. We summarize the most relevant studies below.

**Black-Box Model Extraction.** Early efforts examined the feasibility of stealing shallow or classical models using only prediction APIs (Tramèr et al. 2016). Later work demonstrated that neural NLP systems could also be cloned with modest query budgets: Krishna et al. (2020) reproduced BERT-based translation behavior, and Wallace, Stern, and Song (2020) observed similar vulnerabilities in commercial MT services. Recent studies shift attention to large language models, showing extraction via hard-label outputs (Sha and Zhang 2024) or even partial reconstruction of internal logits (Finlayson, Ren, and Swayamdipta 2024). These advances highlight that LLMs remain susceptible to extraction even when only coarse outputs are available. However, most existing methods focus on general-purpose behaviors or benchmark-oriented tasks. They rarely address how to organize queries in domains with rich terminology, or how to adapt query selection based on surrogate's uncertainty. Complementary work such as ZeroGen (Ye et al. 2022) relies on PLMs to synthesize task data without interacting with the target model; while for data-free model compression, such approaches assume full access to a generator and do not operate in truly restricted black-box environments.

**Query-Efficient Domain Knowledge Stealing.** Another branch of research studies how to use limited model outputs to build effective domain-aware surrogates. Early zero-shot or data-free distillation methods (Wang 2021) emphasized training without labeled datasets. Subsequent works incorporate active querying: Model Leeching (Birch et al. 2023) and Lion (Jiang et al. 2023) show that basic prompt engineering can substantially reduce the number of API calls, while Orca (Mukherjee et al. 2023) highlights the usefulness of extracting explanation traces. EvoKD (Liu et al. 2024) further introduces evolutionary query scheduling to identify weaknesses in the surrogate. These methods collectively demonstrate that the structure and ordering of queries strongly influence distillation quality.

Despite these advances, existing approaches often depend on manually crafted prompts, rely on external domain corpora, or require repeated reinforcement-style optimization. Few methods explicitly leverage the hierarchical structure of domain concepts or adapt follow-up queries based on semantic inconsistencies between the surrogate and the target. Our work differs by initializing the process with structured, concept-grounded templates and then expanding the query space through uncertainty- and feedback-driven refinement. This design enables efficient exploration while operating under strict black-box and data-free constraints.

## Conclusion

We presented an automated framework for extracting domain-relevant knowledge from proprietary large language models through black-box APIs. The method generates semantically grounded prompts from a small set of initial concepts and expands them using uncertainty-driven follow-up queries and diversity-guided exploration, enabling more efficient coverage of domain topics. These findings suggest that effective query design plays an important role in model extraction and underline the need for stronger safeguards for deployed LLMs.

## Discussion

Our framework improves accuracy and query efficiency, but its evaluation is limited to a small set of classification-style tasks, which cannot represent broader domain reasoning or long-form abilities; future work will extend testing to more complex domains. This study also highlights a security risk: all experiments followed standard API rules, and understanding how structured querying accelerates knowledge extraction may help providers design stronger defenses such as rate limiting, monitoring, or response perturbation.

## Acknowledgments

We thank our advisor and colleagues for their helpful discussions and feedback during the development of this work and also appreciate the anonymous reviewers constructive comments, which helped improve the clarity of this paper.

## References


Abdali, S.; He, J.; Barberan, C.; and Anarfi, R. 2024. Can llms be fooled? investigating vulnerabilities in llms. *arXiv preprint arXiv:2407.20529*.

Aguilera-Martínez, F.; and Berzal, F. 2025. LLM Security: Vulnerabilities, Attacks, Defenses, and Countermeasures. *arXiv preprint arXiv:2505.01177*.

Biggs, J. B.; and Collis, K. F. 1982. *Evaluating the Quality of Learning: The SOLO Taxonomy (Structure of the Observed Learning Outcome)*. New York: Academic Press. ISBN 978-0120975501.

Birch, L.; Hackett, W.; Trawicki, S.; Suri, N.; and Garraghan, P. 2023. Model Leeching: An Extraction Attack Targeting LLMs. arXiv:2309.10544.

Bloom, B. S. 1956. *Taxonomy of Educational Objectives, Handbook 1: Cognitive Domain*. Reading, MA: Addison-Wesley Publishing Company.

Carlini, N.; Ippolito, D.; Jagielski, M.; Lee, K.; Tramer, F.; and Zhang, C. 2023. Quantifying Memorization Across Neural Language Models. In *The Eleventh International Conference on Learning Representations*.

Craswell, N.; Mitra, B.; Yilmaz, E.; Campos, D.; and Voorhees, E. M. 2020. Overview of the TREC 2019 deep learning track. arXiv:2003.07820.

Finlayson, M.; Ren, X.; and Swayamdipta, S. 2024. Logits of API-Protected LLMs Leak Proprietary Information. arXiv:2403.09539.

He, J.; Hou, G.; Jia, X.; Chen, Y.; Liao, W.; Zhou, Y.; and Zhou, R. 2024. Data stealing attacks against large language models via backdooring. *Electronics*, 13(14): 2858.

Hosanagar, K.; and Krishnan, R. 2024. Who Profits the Most From Generative AI? *MIT Sloan Management Review*, 65(3): 24–29.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685.

Jeong, C. 2023. A study on the implementation of generative ai services using an enterprise data-based llm application architecture. *arXiv preprint arXiv:2309.01105*.

Jiang, Y.; Chan, C.; Chen, M.; and Wang, W. 2023. Lion: Adversarial Distillation of Proprietary Large Language Models. arXiv:2305.12870.

Jin, D.; Pan, E.; Oufattole, N.; Weng, W.; Fang, H.; and Szolovits, P. 2020. What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams. *arXiv preprint arXiv:2009.13081*.

Jin, Q.; Dhingra, B.; Liu, Z.; Cohen, W. W.; and Lu, X. 2019. 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)*, 2567–2577.

Kringelum, J. V.; Kjærulff, S.; Brunak, S.; Lund, O.; Oprea, T. I.; and Taboureau, O. 2016. ChemProt-3.0: a global chemical biology diseases mapping. *Database: The Journal of Biological Databases and Curation*, 2016: bav123.

Krishna, K.; Tomar, G. S.; Parikh, A. P.; Papernot, N.; and Iyyer, M. 2020. Thieves on Sesame Street! Model Extraction of BERT-based APIs. arXiv:1910.12366.

La Malfa, E.; Petrov, A.; Frieder, S.; Weinhuber, C.; Burnell, R.; Nazar, R.; Cohn, A.; Shadbolt, N.; and Wooldridge, M. 2024. Language-models-as-a-service: Overview of a new paradigm and its challenges. *Journal of Artificial Intelligence Research*, 80: 1497–1523.

Li, X.; and Roth, D. 2002. Learning Question Classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Lin, R. Y.; Ojha, S.; Cai, K.; and Chen, M. F. 2024. Strategic collusion of LLM agents: Market division in multi-commodity competitions. *arXiv preprint arXiv:2410.00031*.

Liu, C.; Zhao, F.; Kuang, K.; Kang, Y.; Jiang, Z.; Sun, C.; and Wu, F. 2024. Evolving Knowledge Distillation with Large Language Models and Active Learning. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 6717–6731. Torino, Italia: ELRA and ICCL.

LW, A.; DR, K.; PW, A.; KA, C.; Mayer, R.; PR, P.; Raths, J.; and MC, W. 2001. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. ISBN ISBN: 080131903X.

Malo, P.; Sinha, A.; Korhonen, P. J.; Wallenius, J.; and Takala, P. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4): 782–796.

Mukherjee, S.; Mitra, A.; Jawahar, G.; Agarwal, S.; Palangi, H.; and Awadallah, A. 2023. Orca: Progressive Learning from Complex Explanation Traces of GPT-4. arXiv:2306.02707.

Orekondy, T.; Schiele, B.; and Fritz, M. 2019. Knockoff nets: Stealing functionality of black-box models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4954–4963.

Pustejovsky, J. 1995. *The Generative Lexicon*. MIT Press.

Ran, K.; Alaofi, M.; Sanderson, M.; and Spina, D. 2025. Two Heads Are Better Than One: Improving Search Effectiveness Through LLM-Generated Query Variants. In *Proceedings of the 2025 ACM SIGIR Conference on Human Information Interaction and Retrieval*, 333–341.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv:1908.10084.

Sha, Z.; and Zhang, Y. 2024. Prompt Stealing Attacks Against Large Language Models. arXiv:2402.12959.

Shah, A.; Paturi, S.; and Chava, S. 2023. Trillion Dollar Words: A New Financial Dataset, Task & Market Analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*.

Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Hou, L.; Clark, K.; Pfohl, S.; Cole-Lewis, H.; Neal, D.;

Schaekermann, M.; Wang, A.; Amin, M.; Lachgar, S.; Mansfield, P.; Prakash, S.; Green, B.; Dominowska, E.; y Arcas, B. A.; Tomasev, N.; Liu, Y.; Wong, R.; Semturs, C.; Mahdavi, S. S.; Barral, J.; Webster, D.; Corrado, G. S.; Matias, Y.; Azizi, S.; Karthikesalingam, A.; and Natarajan, V. 2023. Towards Expert-Level Medical Question Answering with Large Language Models. arXiv:2305.09617.

Sinha, A.; and Khandait, T. 2020. Impact of News on the Commodity Market: Dataset and Results. arXiv:2009.04202.

Smith, B.; Arp, R.; and Spear, A. 2015. *Building Ontologies with Basic Formal Ontology*. Cambridge, MA: MIT Press. ISBN 978-0-262-52781-1. A comprehensive MIT Press textbook on BFO.

Sun, X.; Cheng, G.; Li, H.; Lang, C.; and Han, J. 2024. Stdatav2: Accessing efficient black-box stealing for adversarial attacks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Tramèr, F.; Zhang, F.; Juels, A.; Reiter, M. K.; and Ristenpart, T. 2016. Stealing Machine Learning Models via Prediction APIs. arXiv:1609.02943.

Truong, J.-B.; Maini, P.; Walls, R. J.; and Papernot, N. 2021. Data-free model extraction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4771–4780.

Wallace, E.; Stern, M.; and Song, D. 2020. Imitation Attacks and Defenses for Black-box Machine Translation Systems. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 5531–5546. Online: Association for Computational Linguistics.

Wang, Z. 2021. Zero-Shot Knowledge Distillation from a Decision-Based Black-Box Model. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 10675–10685. PMLR.

Wu, Q.; Bansal, G.; Zhang, J.; Wu, Y.; Li, B.; Zhu, E.; Jiang, L.; Zhang, X.; Zhang, S.; Liu, J.; et al. 2024. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*.

Wu, S.; Irsoy, O.; Lu, S.; Dabravolski, V.; Dredze, M.; Gehrmann, S.; Kambadur, P.; Rosenberg, D.; and Mann, G. 2023. BloombergGPT: A Large Language Model for Finance. arXiv:2303.17564.

Xu, M.; Liao, J.; Wu, J.; He, Y.; Ye, K.; and Xu, C. 2025. Cloud Native System for LLM Inference Serving. *arXiv preprint arXiv:2507.18007*.

Ye, J.; Gao, J.; Li, Q.; Xu, H.; Feng, J.; Wu, Z.; Yu, T.; and Kong, L. 2022. ZeroGen: Efficient Zero-shot Learning via Dataset Generation. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 11653–11669. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Yin, W.; Xu, M.; Li, Y.; and Liu, X. 2024. Llm as a system service on mobile devices. *arXiv preprint arXiv:2403.11805*.

Zhang, M.; Yuan, B.; Li, H.; and Xu, K. 2024. LLM-Cloud Complete: Leveraging cloud computing for efficient large language model-based code completion. *Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023*, 5(1): 295–326.

Zhuang, Y.; Shi, C.; Zhang, M.; Chen, J.; Lyu, L.; Zhou, P.; and Sun, L. 2024. Unveiling the Secrets without Data: Can Graph Neural Networks Be Exploited through {Data-Free} Model Extraction Attacks? In *33rd USENIX Security Symposium (USENIX Security 24)*, 5251–5268.