Active Domain Knowledge Acquisition with $100 Budget: Enhancing LLMs via Cost-Efficient, Expert-Involved Interaction in Sensitive Domains

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have demonstrated an impressive level of general knowledge. However, they often struggle in highly specialized and sensitive domains such as drug discovery and rare disease research due to the lack of expert knowledge, which is often costly to obtain. In this paper, we propose a novel framework (PU-ADKA) designed to efficiently enhance domain-specific LLMs by actively engaging domain experts within a fixed budget. Unlike traditional fine-tuning approaches, PU-ADKA proactively identifies and queries the most appropriate expert from a team, taking into account each expert's availability, competency, knowledge boundaries, and consultation cost. We train PU-ADKA using simulations on PubMed publication data and validate it through domain expert interactions, showing promising improvements in LLM domain knowledge acquisition. Furthermore, our experiments with a real-world drug development team validate that PU-ADKA can significantly enhance LLM performance in specialized domains while adhering to strict budget constraints. In addition to outlining our methodological innovations and experimental results, we release a new benchmark dataset, CKAD, for cost-effective LLM domain knowledge acquisition to foster further research in this challenging area.

1 Introduction

Recent advancements in large language models (LLMs) have led to impressive performance gains across a wide range of tasks (Naveed et al., 2023; Thirunavukarasu et al., 2023; Wu et al., 2024a). However, these gains are not uniformly observed across all domains. In highly specialized, private and sensitive fields, such as drug discovery and rare disease exploration, the acquisition of domain knowledge remains a challenge. Traditional approaches like Reinforcement Learning from Human Feedback (RLHF) (Kaufmann et al., 2023)

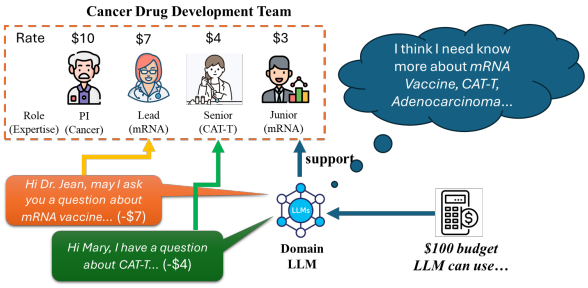

Figure 1: Domain LLM Knowledge Acquisition via Cost-Efficient, Expert-Involved Interaction

have demonstrated value in general settings, yet they struggle in contexts where expert knowledge is extremely expensive and sparse. This scenario is particularly pronounced in domains where domain expertise is fragmented among professionals with diverse competencies and availability constraints (Cheetham and Chivers, 2005). Consequently, there is a pressing need for novel approaches that can efficiently integrate domain expert feedback into LLMs while operating under tight budgetary and expert availability restrictions.

To respond to this demand, we propose the Positive Unlabeled Active Domain Knowledge Acquisition (PU-ADKA), which is designed to proactively engage with domain experts and selectively acquire targeted feedback that can significantly enhance the performance of LLMs in specialized fields. Unlike conventional fine-tuning methods that passively incorporate affordable human feedback (Zhang et al., 2023), PU-ADKA actively queries the most appropriate expert from a team given each member's computational profile. The model can elaborately consider factors such as the candidate expert's knowledge boundary, cost of consultation, and expert availability, thereby optimizing the knowledge acquisition process within a fixed budget (e.g., total $100). The model training process leveraged newly released domain knowledge (e.g., recent PubMed data (White, 2020)),

legacy architectures of LLMs and innovative simulations of expert-domain knowledge interactions. Through an intelligent knowledge selection process and cost-aware querying mechanism, PU-ADKA bridges the gap between the limited availability of expert input and the high demand for domain-specific information.

Figure 1 illustrates the concept behind the proposed PU-ADKA. In this case, a domain LLM acknowledges gaps in its knowledge related to topics like *mRNA vaccines*, *CAT-T*, and *adenocarcinoma* (to support a cancer drug development team) (Kalyuga, 2007). Instead of relying on static, pre-existing datasets, PU-ADKA proactively engages with domain experts to acquire precise knowledge within a limited budget. The model evaluates the expertise, cost, and availability of different specialists, including PI, lead, senior, and junior scholars, to optimize knowledge acquisition. For example, in the image, the LLM selectively queries Dr. Jean for insights on *mRNA vaccines* at a cost of $7, while consulting Mary, a different expert, about *CAT-T* for $4, ensuring cost-effective expert engagement. This dynamic querying mechanism allows the LLM to refine its domain knowledge efficiently, making it particularly useful in critical domains like drug discovery and rare disease research, where expert knowledge is both sparse and expensive.

The contribution of this paper is fourfold:

- **Methodology**: We introduce PU-ADKA, a proactive, cost-efficient model that strategically queries domain experts to enhance LLM performance in highly specialized fields with very limited expert availability.
- **Cost-Aware Expert Selection**: We develop a mechanism that considers expert competency, knowledge boundaries, availability, and consultation cost, ensuring that each query yields maximum value under a fixed budget.
- **Experiment**: We validate the efficacy of PU-ADKA using both simulation evaluation and real-world cancer drug development study. The latter experiment used a real drug development team where five experts with diverse background participate in the experiment. The result shows that PU-ADKA is promising to enhance domain LLMs with a fixed budgetary restriction.
- **Benchmark Dataset**: To foster further research in the area of domain-specific LLM enhancement, we provide a new benchmark dataset, Cost-Aware Knowledge Acquisition Dataset (CKAD), for LLM domain knowledge acquisition, which is available for open access.

2 Related Work

2.1 Human Feedback Integration in Domain-Specific LLMs

Domain-specific adaptation of LLMs has been advanced significantly by techniques such as domain-adaptive pretraining (DAPT) (Gururangan et al., 2020) and various biomedical LLMs like BioMedLM (Bolton et al., 2024), ClinicalBLIP (Ji et al., 2024), and BioGPT (Luo et al., 2022). These methods effectively utilize large domain-specific corpora (e.g., PubMed) to incorporate static knowledge. However, they often fall short in capturing the dynamic insights from domain experts, crucial for rapidly evolving areas like drug discovery. Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022) aims to align general LLMs with human preferences but typically depends on more homogeneous and less costly annotators, limiting its effectiveness in specialized domains where expert feedback is sparse and expensive. Attempts like ExpertQA (Malaviya et al., 2023) simulate multi-expert interactions but overlook practical constraints like budget limitations and asynchronous availability of experts. Our approach, ADKAM, overcomes these shortcomings by redefining expert knowledge acquisition as a budget-constrained optimization task, selectively engaging experts based on their competence, cost, and availability, thereby transitioning from static data-driven adaptation to proactive, expert-guided learning.

2.2 Budget-Constrained Active Learning with Multi-Expert Collaboration

Traditional active learning models primarily focus on maximizing sample information through uncertainty (Gal et al., 2017; Kim et al., 2021) or diversity (Chakraborty et al., 2015; Parvaneh et al., 2022; Citovsky et al., 2021), often neglecting the varying costs associated with expert annotations, particularly in complex fields like biomedicine. Cost-sensitive approaches (Huang et al., 2017; Henkel et al., 2023) attempt to address this by optimizing for lower-cost annotators but fail to differentiate

between the varied expertise levels necessary for accurately labeling complex cases. Unlike these methods, ADKAM integrates active learning with strategic expert collaboration, emphasizing both data sample selection based on potential to update the model and efficient engagement of experts, balancing cost against their competency and availability.

2.3 Sampling Strategy in LLM Active Learning

The importance of data instances in modern LLMs is often evaluated through gradient-based (Xia et al., 2024; Wu et al., 2024b), similarity-based (Xie et al., 2023; Li et al., 2023a), or in-context learning (Li et al., 2023a) methods. These techniques typically assume samples are independent and identically distributed, a premise that does not hold in complex fields with interdependent data such as biomedical texts. Despite progress in sampling strategies that account for diversity (Liu et al., 2023), many approaches do not consider the costs associated with expert annotations or the specific expertise required for accurate data labeling. Our proposed method, PU-ADKA, addresses these challenges by prioritizing high-impact samples through a refined uncertainty estimation specific to the domain and strategically assigning these samples to the most cost-effective experts capable of providing high-quality annotations. This method ensures that knowledge acquisition is not only efficient but also economically feasible in constrained environments.

3 Methodology

In this section, we begin by formalizing the cost-aware LLM knowledge acquisition problem (Section 3.1). We then present PU-ADKA framework for efficient domain-specific LLM knowledge acquisition in Figure 2. PU-ADKA addresses two key challenges: (1) How to leverage LLMs to simulate active learning in high-cost domains? (Section 3.2) and (2) How to simultaneously optimize data selection and cost-aware expert assignment for maximal knowledge acquisition under fixed budgets? (Section 3.3 and Section 3.4)

3.1 Problem Definition

Given a fixed annotation budget B, an unlabeled question pool $\mathcal{D}_{\text{tr}} = \{q_i\}_{i=1}^{|\mathcal{D}_{\text{tr}}|}$, and a team of domain experts $\mathcal{E} = \{e_j\}_{j=1}^{|\mathcal{E}|}$, our goal is to select an optimal set of (q_i, e_j) pairs to finetune a large language model θ, maximizing finetuning performance on a target test set $\mathcal{D}_{\text{te}} = \{p_m\}_{m=1}^{|\mathcal{D}_{\text{te}}|}$.

Formally, we define an allocation function $f : \mathcal{D}_{\text{tr}} \to \mathcal{E}$ that assigns each selected question q_i to an expert e_j, ensuring that the total annotation cost remains within the budget B. The optimization objective is:

$$\mathcal{S}^* = \arg\max_{\mathcal{S} \subseteq \mathcal{D}_{\text{tr}} \times \mathcal{E}} \mathcal{F}(\theta_{\mathcal{S}}, \mathcal{D}_{\text{te}})$$
$$\text{s.t.,} \sum_{(q_i, e_j) \in \mathcal{S}} c(q_i, e_j) \leq B,$$

where, $\mathcal{S}^*$ denotes the optimal set of (q_i, e_j) pairs that maximizes the performance metric $\mathcal{F}(\theta_{\mathcal{S}}, \mathcal{D}_{\text{te}})$ of the fine-tuned model $\theta_{\mathcal{S}}$ on the target test set. The term $c(q_i, e_j)$ represents the annotation cost incurred when expert e_j annotates question q_i.

3.2 Simulation Environment Construction

To support our investigation, we pioneered a new benchmark data, Cost-Effective Knowledge Acquisition Dataset (CKAD), for simulating biomedical expert consultations and LLM knowledge acquisition process by strategically leveraging the comprehensive knowledge within the PubMed digital library. This approach harnesses previously untapped domain expertise and research findings, predating the knowledge cutoff of selected LLMs, to construct robust datasets for consultation simulation and model training/evaluation. To simulate the knowledge acquisition process, we introduce a temporal knowledge separation method based on PubMed data, which ensures strict chronological isolation between the base model's pre-existing knowledge and newly acquired target domain knowledge through three core components.

Predated Base Model Selection: We employ Llama2-7B (Touvron et al., 2023) as our predated base model, chosen for its knowledge limitations to information available up to early 2023, prior to our target corpus. This temporal separation ensures a controlled setting for evaluating knowledge acquisition.

Temporal Corpus Construction: We construct CKAD from 2024 PubMed Central (PMC) (Fiorini et al., 2017), extracting question-answer (QA) pairs using GPT-4o (OpenAI, 2024). For each paper, five mechanism-focused QA pairs are generated using prompting and manually validated. To maintain a clean environment for assessing knowledge acquisi-

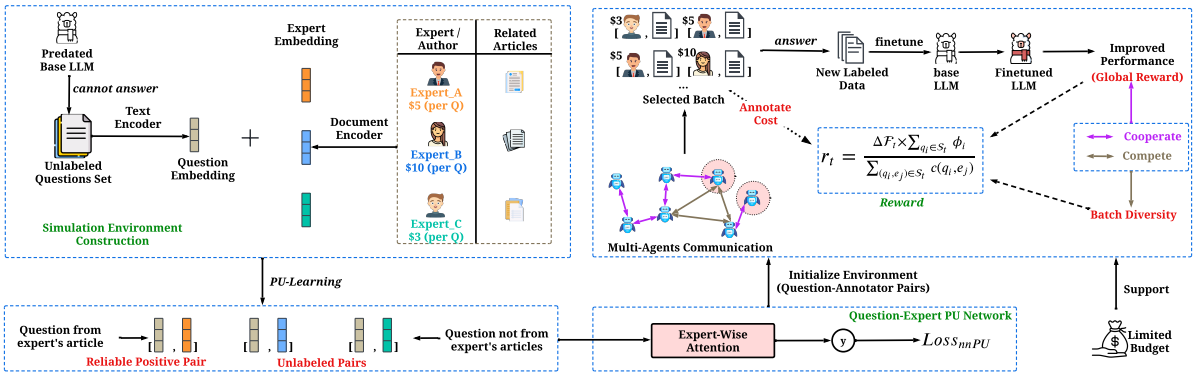

Figure 2: Illustration of our proposed PU-ADKA framework.

tion, we filter out QA pairs that can be answered by the base model, using GPT-4o as the judge model[1]. This process results in a final dataset of *48,219 QA pairs* (the base model cannot correctly answer) representing post-2023 knowledge.

Expert Simulation: We model real-world annotation constraints by assigning expert roles to the top 20 authors ranked by publication count, using them as proxy experts. GPT-4o is deployed to estimate a binary expert capability matrix $A \in \mathbb{R}^{Q \times N}$ where each entry A_{ji} is set to 1 if expert e_j is capable of labeling question q_i, and 0 otherwise. The matrix is constructed by leveraging relevant sections of experts' papers to assess their expertise. To determine a reasonable unit labeling price, we rank experts based on the sum of their papers' impact factors (Clarivate, 2025) and assign higher rates to those with higher ranks in a principled manner.

This approach maintains chronological separation by integrating a constrained base model, a curated QA dataset, and expertise-driven annotation, preventing knowledge leakage while maintaining a controlled knowledge acquisition setting.

3.3 Expert Profiling with PU Learning

We formalize the question-expert matching task as a Positive-Unlabeled (PU) learning problem, which helps to characterize each expert's knowledge boundary. Given a question-expert pair (q_i, e_j), we label it as ***positive*** if q_i originates from a publication authored by e_j. However, if q_i does not come from e_j's paper, we do not automatically treat (q_i, e_j) as a negative pair. Instead, it remains ***unlabeled*** because the expert may be qualified to label the question. For instance, a scholar specializing in *cancer NK cells* could potentially annotate a *sepsis-related question* if they possess relevant medical expertise (e.g., *extracellular vesicles*), even without publications on *sepsis*.

We use LLM-based text representations, leveraging a pretrained Llama2-7B model to encode questions E_q^i and experts E_e^j, with embeddings taken from the last hidden layer. Particularly, an expert's embedding is obtained by averaging the representations of their publications (Wu et al., 2023). To train our PU model to estimate expert knowledge boundary, we employ an expert-wise attention mechanism [2] and training with the non-negative PU risk estimator (Kiryo et al., 2017), which is defined as follows:

$$\begin{aligned} \text{Risk}_{pu}(g) = & \frac{\pi_p}{n_p}\sum_{i=1}^{n_p} l(g(x_k^p), +1) + \\ & max(0, \frac{1}{n_u}\sum_{i=1}^{n_u} l(g(x_k^u), -1) - \\ & \frac{\pi_p}{n_p}\sum_{i=1}^{n_p} l(g(x_k^p), -1)), \end{aligned} \quad (1)$$

where π denotes positive class prior ($\pi = 0.1$ in our dataset), $l(\cdot, \cdot)$ is the surrogate loss of zero-one loss (Du Plessis et al., 2015), n_p represents the number of labeled positive instances, n_u represents the number of unlabeled instances, x_k^p and x_k^u denote question-annotator pair in the labeled positive set and the unlabeled set, respectively.

3.4 Efficient Domain Knowledge Acquisition via Multi-Agent Reinforcement Learning

Given the budgetary constraints, we proposed a novel model that aligns question selection with available expert knowledge while maximizing the LLM's domain knowledge acquisition competency.

[1]The details of question-answer extraction and the evaluation prompt are provided in the Appendix A.

[2]The attention network is detailed in Appendix **??**

To ensure that the selected question set captures both informativeness and diversity, we formulate the selection process as a multi-agent reinforcement learning (RL) problem, where each agent is tasked with selecting a (question-expert) pair. The number of agents, n, determines the size of the question-set at each iteration. Unlike traditional RL models, the proposed interactive multi-agent RL can estimate the sampled question-expert pair importance by leveraging inter-agent competition and cooperation, ensuring both informational density and diversity in selected pairs with a fixed budget.

3.4.1 Multi-Agent RL State

The environment state is represented by a combination of features that capture both task-related and budgetary aspects: (1). The question–expert matching score $g(q_i, e_j)$ is derived from the trained PU learning model and measures the suitability of assigning question q_i to expert e_j. (2). The remaining budget B_t indicates the available annotation budget at time step t. (3). The expert sampling probability quantifies the likelihood of selecting each expert e_j, defined as:

$$w_j^t = \frac{B_t}{c(q_i, e_j)} \times (1 - \alpha \Gamma_j^t), \tag{2}$$

where α is a decay factor, and Γ_j^t denotes the number of times expert e_j has been selected up to time step t. This formulation encourages diversity in expert selection to enhance overall information gain while ensuring balanced workload distribution.

3.4.2 Multi-Agent Communication

Competition. Different from previous studies, our framework allows multiple agents within the same model to simultaneously seek (q_i, e_j) pairs, enabling different experts to compete for answering the same question. Leveraging our PU-based question-expert matching model, each question q_i is associated with a ranked list of potential experts. As a result, multiple experts $e_1, e_2, \ldots, e_h$ may select the same question q_i. In such cases, q_i should be assigned to the expert with the highest matching score based on our PU matching network. To enforce this competitive selection, we introduce a competition function:

$$\begin{aligned} &Compete(q_i \mid e_1, e_2, \ldots, e_h) = e_z, \\ &\text{s.t.} \quad e_z = \arg\max_{e_j} g(q_i, e_j), \end{aligned} \tag{3}$$

where $g(q_i, e_j)$ represents the PU-based matching score between question q_i and expert e_j, ensuring that the most suitable expert is selected. For experts who lose the competition for a given question in the current iteration, the corresponding agents will then select alternative pairs and re-enter the competition process. This recursive procedure continues until all agents in the current state have been assigned unique questions.

Cooperation. To effectively encourage collaborative decision-making among agents and optimize knowledge acquisition under a fixed annotation budget, we define the reward function as:

$$r_t = \frac{\Delta \mathcal{F}_t \times \sum_{q_i \in \mathcal{S}_t} \phi_i}{\sum_{(q_i, e_j) \in \mathcal{S}_t} c(q_i, e_j)}, \tag{4}$$

where $\Delta \mathcal{F}_t$ denotes the improvement in model performance on the validation set after incorporating newly labeled data at step t, and the denominator represents the total annotation cost (Gao and Saar-Tsechansky, 2020; Huang et al., 2017; Golazizian et al., 2024). The diversity term ϕ_i measures the distinctiveness of each selected question and is defined as:

$$\phi_i = \min_{q_z \in \mathcal{S}_t} d(E_q^i, E_q^z), \tag{5}$$

where $\mathcal{S}_t$ denotes the cureent labeled question set, and $d(\cdot, \cdot)$ is the Euclidean distance function. A larger ϕ_i value indicates that the selected question is more diverse relative to past selections, thereby enhancing knowledge coverage and reducing redundancy.

3.4.3 Model Training

To stabilize learning, we employ a Double DQN architecture (Wang et al., 2020). The temporal-difference (TD) target is computed as:

$$Y_t = r_t + \gamma Q(s_{t+1}, \arg\max_{u_{t+1}} Q(s_{t+1}, u_{t+1}; \theta_t); \theta_t'), \tag{6}$$

where s_{t+1} denotes the next state, γ is the discount factor, θ_t and θ_t' represent to the parameters of the policy and target network, respectively. To enhance generalization, we employ bootstrap sampling by selecting a random subset of experts (e.g. five per iteration) during training stage. This strategy prevents overfitting to a specific set of experts, ensuring that the learned policy remains robust across diverse labeling scenarios.

4 Experiments

4.1 Experimental Settings

Model Architecture and Training Settings. As described in Section 3.2, we use the PubMed dataset for sepsis and cancer NK research from 2024 and adopt Llama2-7B as the base architecture. The experimental setup for our PU-ADKA model utilizes Llama2-7B with a sampling temperature of 1.0, a nucleus sampling top_p value of 0.9, and a maximum token length of 4,096. The question and expert document encoders use the last hidden layer of Llama2-7B. For fine-tuning, we apply LoRA (Hu et al., 2021) to improve training efficiency for large-scale models. The LoRA configuration includes a rank of 16, an alpha of 128, and a dropout rate of 0.1. Training involves learning LoRA matrices for all attention mechanisms in each configuration. The models are optimized using the AdamW optimizer with a learning rate of 2×10^{-5}. Each configuration undergoes three trials with different random seeds.

Table 1: Statistics of CKAD dataset.

Disease Type	Cancer_NK and Sepsis
#Train	38,575
#Dev	4,722
#Test	4,722

In the multi-agent reinforcement learning framework, we employ the Double DQN (Wang et al., 2020) architecture. The default number of agents is 10, with five experts selected per iteration. In each iteration, experts are ranked based on the sum of their papers' impact factors (Clarivate, 2025), and their unit prices are assigned accordingly as $[\$0.5, \$0.4, \$0.3, \$0.2, \$0.1]$ per labeled question. The total annotation budget is set to 100.

Evaluation Benchmarks and Metrics. To ensure a clean evaluation of knowledge acquisition, our experimental dataset consists of general disease mechanism question-answer pairs that cannot be answered initially (i.e., the initial answerable rate is 0). Details of the dataset are provided in Table 1. During the simulation training stage, we employ two advanced models, GPT-4o-2024-08-06 and GPT-4-Turbo, as judge models. The evaluation metrics include win rate and length-controlled win rate.

Additionally, we conduct human-involved experiments to validate the effectiveness of our method. Our expert team consists of three sepsis specialists and two cancer specialists, representing different levels of expertise. Among them, one is a principal investigator (PI), while the remaining members include one medical doctor and three PhD students.

Baselines. To ensure a comprehensive evaluation, our experiment includes a variety of baseline methodologies that encompass both question selection and expert allocation strategies. The comparison provides insights into the effectiveness of different active learning frameworks applied to LLMs. Below we detail the baselines used:

- **Random:** Questions are selected randomly, providing a baseline for minimal strategic intervention in data selection.
- **DEITA:** (Liu et al., 2023) Evaluates data across complexity, quality, and diversity using pretrained complexity scorer[3] and quality scorer [4] to score each unlabeled questoions.
- **CHERRY:**(Li et al., 2023a) Applies the Instruction-Following Difficulty (IFD) metric to assess question quality autonomously.
- **NUGGETS:**(Li et al., 2023b) Assesses the relevance of questions by considering each as a single instance in one-shot learning contexts.
- **LESS:** (Xia et al., 2024) Calculates the influence of questions on the validation set to prioritize data that may yield the most significant insights during finetuning.
- **ROSE:**(Wu et al., 2024b) Utilizes gradient similarity to evaluate the potential contribution of each question to the model's performance, aligning with active learning principles of uncertainty and diversity.

For expert allocation, we implement the following methods:

- **Random:** Experts are assigned randomly to questions.
- **Cost-Greedy:** This method always selects the least expensive expert available, optimizing for cost efficiency.

[3]https://huggingface.co/hkust-nlp/deita-complexity-scorer

[4]https://huggingface.co/hkust-nlp/deita-complexity-scorer

Table 2: Overall Performance Comparison on CKAD dataset (%).

Category	Model	GPT-4o-2024-08-06	GPT-4-Turbo	GPT-4o-2024-08-06	GPT-4-Turbo	Avg.Length
		WR	WR	LC_WR	LC_WR	-
Random	RAND	4.7 (0.4)	6.7 (0.8)	20.3 (0.9)	20.4 (0.8)	2220
	DEITA	9.6 (0.3)	7.9 (0.1)	21.0 (0.9)	22.1 (0.8)	2212
	CHERRY	7.8 (0.1)	8.3 (0.2)	20.4 (0.9)	21.5 (0.9)	2221
	NUGGETS	10.4 (0.1)	10.7 (0.4)	21.0 (0.8)	20.4 (0.8)	2204
	LESS	7.9 (0.2)	7.9 (0.2)	22.0 (1.0)	24.0 (1.1)	2212
	ROSE	8.1 (0.4)	10.0 (0.2)	21.5 (1.0)	22.7 (1.0)	2194
Cost-Greedy	RAND	6.2 (0.4)	6.7 (0.8)	20.4 (0.9)	20.5 (0.9)	2207
	DEITA	14.2 (0.8)	11.7 (0.2)	20.9 (1.0)	20.9 (0.9)	2246
	CHERRY	11.7 (0.3)	10.0 (0.4)	23.4 (0.9)	22.1 (1.1)	2236
	NUGGETS	7.9 (0.4)	8.7 (0.4)	21.5 (0.9)	20.4 (0.9)	2182
	LESS	12.1 (0.4)	9.6 (0.4)	22.1 (0.8)	21.2 (1.0)	2218
	ROSE	8.3 (0.8)	9.7 (0.2)	20.4 (0.9)	22.7 (1.0)	2174
Match-Greedy	RAND	6.7 (0.8)	7.9 (0.4)	20.9 (1.0)	19.9 (0.8)	2204
	DEITA	10.0 (0.3)	9.2 (0.8)	21.2 (1.0)	22.3 (0.9)	2214
	CHERRY	7.5 (0.0)	9.2 (0.2)	21.0 (0.9)	23.3 (1.1)	2173
	NUGGETS	9.5 (0.3)	11.6 (0.2)	22.1 (1.0)	21.6 (0.9)	2182
	LESS	12.1 (0.4)	10.4 (0.2)	23.5 (1.0)	22.5 (1.0)	2252
	ROSE	9.2 (0.1)	10.9 (0.4)	22.5 (0.9)	21.9 (1.0)	2229
Ours	PU-ADKA	**18.2** (0.6)	**16.7** (0.4)	**25.6** (1.0)	**26.5** (0.9)	1781

- **Match-Greedy:** Matches questions to experts based on the highest embedding similarity between them, facilitating a more informed allocation.

These baselines are integral to understanding the landscape of active learning strategies within LLM contexts, providing a benchmark against which our proposed methods can be evaluated.

4.2 Experimental Results

Experimental Results Our experimental results are detailed in Table 2, where we compare the performance of our method, PU-ADKA, against various baseline strategies. PU-ADKA consistently outperforms all baselines in terms of knowledge acquisition across different judging models. Specifically, with the GPT-40-2024-08-06 model as judge, PU-ADKA achieves a win rate of 18.2% and an LC-WR of 25.65%. When evaluated by the GPT-4-Turbo model, it records a win rate of 16.7% and an LC-WR of 26.57%. These results exceed those of the next best baseline, DEITA under the Cost-Greedy strategy, by margins of 4% and 5% in win rate, and 2.1% and 3.2% in LC-WR, respectively, under the two judging conditions. Noteably, LESS performs stable when under both Cost-Greedy and Match-Greedy settings, the GPT-4o-2024-08-06 and GPT-4-Turbo judge the win rate at 12.1% and 10% in both settings. Furthermore, the minimal baseline performance under fully random conditions, with win rates of 4.7% and 6.7%, highlights the baseline challenge and emphasizes the robustness of our method against less strategic approaches.

Table 3: Human-involved results judged by GPT-4-Turbo.

	WR	LC_WR
Random (Random)	7.5 (0.7)	20.3 (0.8)
LESS (Random)	9.2 (0.5)	20.5(0.9)
LESS (Cost-Greedy)	11.4 (0.6)	21.0 (1.0)
LESS (Match-Greedy)	12.5 (0.7)	21.2 (0.8)
PU-ADKA	**15.2** (0.8)	**24.3** (0.9)

4.3 Human Involved Validation

To further substantiate the robustness of our method, PU-ADKA, we implemented it within a professional biomedical team of experts under a simulated budget constraint of $100 per game. The cost of annotator expertise was varied, reflecting their respective professional knowledge in the domain, with unit prices set at ($0.5, $0.2, $0.1, $0.1, $0.1 per labeling question. We assessed the performance in terms of win rate and LC win rate using GPT-4-Turbo as the judge under various settings: fully random, and LESS for question selection combined with each of the three expert allocation strategies (Random, Cost-Greedy, and Match-Greedy).

The detailed results are presented in Table 3.

The data reveal that PU-ADKA notably surpasses the most competitive baseline, LESS (Match-Greedy), by margins of 2.7% and 3.1% in win rate and LC win rate, respectively. This enhancement in performance in a practical setting underscores the effectiveness of our method, particularly in scenarios constrained by budget. This real-world application not only validates the utility of PU-ADKA but also establishes it as a formidable approach in the domain of budget-limited active learning.

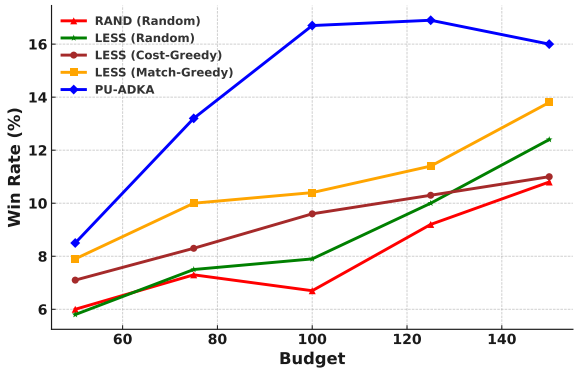

Figure 3: Performance under Different Budgets

Table 4: Ablation results on CKAD dataset with ✓ indicating the enabling of the corresponding module. (Judged by GPT-4-Turbo)

Variant	PU	MA	WR	LC_WR
I		✓	13.3 (0.7)	23.2 (1.1)
II	✓		14.2 (0.6)	23.0 (1.0)
PU-ADKA	✓	✓	**16.7** (0.4)	**26.5** (0.9)

4.4 Ablation Study

4.4.1 Validating the Utility of Each Module

To thoroughly assess the contributions of each component within **PU-ADKA**, specifically the multi-agent (MA) framework and the positive-unlabeled (PU) learning approach, we performed a series of ablation studies. These studies were conducted on the QA dataset, with GPT-4-Turbo serving as the judge. We explored two key variants:

- **Variant I:** Utilizes unsupervised embedding-based similarity measures in place of the PU learning model to understand the impact of the PU approach on the overall performance.
- **Variant II:** Operates under a single-agent setup to evaluate the effectiveness of our multi-agent configuration.

The results, detailed in Table 4, highlight the integral role each module plays in the success of **PU-ADKA**. The comparison with Variant I underscores the superiority of our PU-based question-expert matching technique. Similarly, when contrasted with the single-agent model of Variant II, our multi-agent method demonstrates its enhanced capability in expert allocation strategy, confirming the benefits of our comprehensive framework in active learning scenarios.

4.4.2 Performance under Different Budgets

We evaluated the performance of our model, **PU-ADKA**, against various baseline methods under differing budget scenarios, as depicted in Figure 3. The results indicate that our method achieves consistently robust outcomes across all tested budget levels compared to the baselines. Notably, at a budget of $100, **PU-ADKA** significantly outperforms the next best approach, LESS (Match-Greedy). Beyond this budget point, the rate of knowledge acquisition stabilizes, showing no substantial further increases (Han et al.). This plateau suggests that our method is particularly effective at rapidly acquiring knowledge within constrained budget settings, demonstrating a distinct advantage over competing methods in efficiently utilizing available resources.

5 Conclusion and Future Work

This study introduces PU-ADKA, a novel approach designed to enhance LLMs through active learning in domains where expert feedback is prohibitively costly. Distinct from general active learning models that treat expert input uniformly, PU-ADKA strategically engages experts based on their specialized knowledge, availability, and cost-effectiveness. This targeted approach not only optimizes budget utilization but also significantly improves LLM performance. Validated through rigorous simulations and real-world applications in high-cost domains, PU-ADKA demonstrates a superior method for integrating scarce and valuable expert feedback into LLMs. The release of the CKAD dataset further supports ongoing research into domain-specific LLM enhancements

Limitations

Scalability with Increasing Data and Experts As the number of unlabeled data points and avail-

able experts grows, the scale of PU-ADKA changes significantly. Larger datasets require more efficient selection strategies, while an increasing pool of experts introduces greater complexity in allocation and coordination. Future research should explore more scalable solutions to maintain efficiency as the system scales to real-world, large-scale applications.

Impact of Number of Agents and Computational Constraints The number of agents directly affects the system's performance and computational demands. While PU-ADKA operates within a multi-agent framework, we did not extensively experiment with varying agent numbers due to the high computational cost associated with training and coordination. Additionally, we did not explore different batch sizes or report computational efficiency under varying agent settings. Future work should investigate the trade-offs between agent scalability, computational efficiency, and performance optimization.

Generalizability to Other Domains While this study primarily focuses on biomedical expert interactions, other high-cost domains such as law and finance face similar challenges. Expanding PU-ADKA to these fields and evaluating its adaptability to different datasets and model architectures will be essential for broader applicability.

References

Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, et al. 2024. Biomedlm: A 2.7 b parameter language model trained on biomedical text. *arXiv preprint arXiv:2403.18421*.

Shayok Chakraborty, Vineeth Balasubramanian, Qian Sun, Sethuraman Panchanathan, and Jieping Ye. 2015. Active batch selection via convex relaxations with guaranteed solution bounds. *IEEE transactions on pattern analysis and machine intelligence*, 37(10):1945–1958.

Graham Cheetham and Geoffrey E Chivers. 2005. *Professions, competence and informal learning*. Edward Elgar Publishing.

Gui Citovsky, Giulia DeSalvo, Claudio Gentile, Lazaros Karydas, Anand Rajagopalan, Afshin Rostamizadeh, and Sanjiv Kumar. 2021. Batch active learning at scale. *Advances in Neural Information Processing Systems*, 34:11933–11944.

Clarivate. 2025. Master journal list. Accessed: 2025-01-02.

Marthinus Du Plessis, Gang Niu, and Masashi Sugiyama. 2015. Convex formulation for learning from positive and unlabeled data. In *International conference on machine learning*, pages 1386–1394. PMLR.

Nicolas Fiorini, David J Lipman, and Zhiyong Lu. 2017. Towards pubmed 2.0. *Elife*, 6:e28801.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep bayesian active learning with image data. In *International conference on machine learning*, pages 1183–1192. PMLR.

Ruijiang Gao and Maytal Saar-Tsechansky. 2020. Cost-accuracy aware adaptive labeling for active learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 2569–2576.

Preni Golazizian, Alireza S Ziabari, Ali Omrani, and Morteza Dehghani. 2024. Cost-efficient subjective task annotation and modeling through few-shot annotator adaptation. *arXiv preprint arXiv:2402.14101*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Jindong Han, Hao Liu, Jun Fang, Naiqiang Tan, and Hui Xiong. Automatic instruction data selection for large language models via uncertainty-aware influence maximization. In *THE WEB CONFERENCE 2025*.

Julia Henkel, Genc Hoxha, Gencer Sumbul, Lars Möllenbrok, and Begüm Demir. 2023. Annotation cost efficient active learning for content based image retrieval. In *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*, pages 4994–4997. IEEE.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Sheng-Jun Huang, Jia-Lve Chen, Xin Mu, and Zhi-Hua Zhou. 2017. Cost-effective active learning from diverse labelers. In *IJCAI*, pages 1879–1885.

Jia Ji, Yongshuai Hou, Xinyu Chen, Youcheng Pan, and Yang Xiang. 2024. Vision-language model for generating textual descriptions from clinical images: model development and validation study. *JMIR Formative Research*, 8:e32690.

Slava Kalyuga. 2007. Expertise reversal effect and its implications for learner-tailored instruction. *Educational psychology review*, 19:509–539.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2023. A survey of reinforcement learning from human feedback. *arXiv preprint arXiv:2312.14925*.

Yoon-Yeong Kim, Kyungwoo Song, JoonHo Jang, and Il-Chul Moon. 2021. Lada: Look-ahead data acquisition via augmentation for deep active learning. *Advances in Neural Information Processing Systems*, 34:22919–22930.

Ryuichi Kiryo, Gang Niu, Marthinus C Du Plessis, and Masashi Sugiyama. 2017. Positive-unlabeled learning with non-negative risk estimator. *Advances in neural information processing systems*, 30.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2023a. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. *arXiv preprint arXiv:2308.12032*.

Yunshui Li, Binyuan Hui, Xiaobo Xia, Jiaxi Yang, Min Yang, Lei Zhang, Shuzheng Si, Junhao Liu, Tongliang Liu, Fei Huang, et al. 2023b. One shot learning as instruction data prospector for large language models. *arXiv preprint arXiv:2312.10302*.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2023. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. *arXiv preprint arXiv:2312.15685*.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2023. Expertqa: Expert-curated questions and attributed answers. *arXiv preprint arXiv:2309.07852*.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

OpenAI. 2024. Gpt-4o model card.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Amin Parvaneh, Ehsan Abbasnejad, Damien Teney, Gholamreza Reza Haffari, Anton Van Den Hengel, and Javen Qinfeng Shi. 2022. Active learning by feature mixing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12237–12246.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jianhao Wang, Zhizhou Ren, Terry Liu, Yang Yu, and Chongjie Zhang. 2020. Qplex: Duplex dueling multi-agent q-learning. *arXiv preprint arXiv:2008.01062*.

Jacob White. 2020. Pubmed 2.0. *Medical reference services quarterly*, 39(4):382–387.

Yang Wu, Xurui Li, Xuhong Zhang, Yangyang Kang, Changlong Sun, and Xiaozhong Liu. 2023. Community-based hierarchical positive-unlabeled (pu) model fusion for chronic disease prediction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 2747–2756.

Yang Wu, Chenghao Wang, Ece Gumusel, and Xiaozhong Liu. 2024a. Knowledge-infused legal wisdom: Navigating llm consultation through the lens of diagnostics and positive-unlabeled reinforcement learning. *arXiv preprint arXiv:2406.03600*.

Yang Wu, Huayi Zhang, Yizheng Jiao, Lin Ma, Xiaozhong Liu, Jinhong Yu, Dongyu Zhang, Dezhi Yu, and Wei Xu. 2024b. Rose: A reward-oriented data selection framework for llm task-specific instruction tuning. *arXiv preprint arXiv:2412.00631*.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. Less: Selecting influential data for targeted instruction tuning. *arXiv preprint arXiv:2402.04333*.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy S Liang. 2023. Data selection for language models via importance resampling. *Advances in Neural Information Processing Systems*, 36:34201–34227.

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023. Llmaaa: Making large language models as active annotators. *arXiv preprint arXiv:2310.19596*.

A Prompts

In this section, we present the detailed prompts used for generating Question-Answer data based on PubMed 2024 Sepsis and Cancer NK cell papers, as well as the specific prompts employed for model evaluation across all experiments.

A.1 QA Extraction

PubMed Paper Question-Answer Generation

<|im_start|>system
You are an expert in extracting specific and relevant question-answer pairs from scientific papers. Your task is to generate five QA pairs based on the unique mechanisms or processes described in the provided paper. Focus on extracting detailed mechanisms or processes, avoiding generic or summarization-style questions.

Guidelines:
1. The questions must specifically target mechanisms, processes, or detailed explanations provided in the paper. Focus on "how" or "why" certain processes or mechanisms work according to the paper.
2. Avoid generic or summarization-style questions, such as broad overviews or general statements about findings.
3. Each question should be clear, concise, and specific, addressing a mechanism, interaction, or process described in the paper.
4. The answers must directly explain the mechanism or process, based on specific information from the paper, and be precise and to the point.

Examples:
- Question 1: How does cytokine IL-15 regulate the activation of natural killer cells in the study?
Answer: Cytokine IL-15 regulates natural killer cell activation by binding to its receptor, triggering a signaling cascade that enhances proliferation and cytotoxic activity.
- Question 2: What mechanism underlies the feedback loop described for natural killer cell regulation?
Answer: The feedback loop involves cytokine signaling that stimulates metabolic reprogramming in natural killer cells, which in turn amplifies cytokine production.

<|im_end|>
<|im_start|>user

Below is the content of the paper:
<Insert the paper's abstract, introduction, and methodology here.>
Your task is to generate five QA pairs based on the unique mechanisms or processes described in the provided paper. Focus on extracting detailed mechanisms or processes, avoiding generic or summarization-style questions. The response format should be:
<Question: The generated question>
<Answer: The generated answer>
The generated five QA pairs are:

<|im_end|>

A.2 GPT4 Judge Prompt

Evaluation Prompt

<|im_start|>system
You are a teacher assessing whether a Output (b) correctly covers the core meaning of a Output (a) for a given Question. The Output (b) must fully address the question, just as the Output (a) does. Follow these rules strictly:
Scoring Criteria

1. **Semantic Match**: - The Output (b) must **precisely match** the meaning of the Output (a) without significant divergence. - Output (b) must address the Question in the same way as the Output (a).
2. **Supplementary Information**: - Additional details are allowed **only if they do not conflict** with the Output (a). - Output (b) must not contain any contradictions, factual errors, or misleading information.

Evaluation Process

1. **Key Point Extraction**: - Extract core facts, entities, and logical relationships from the Output (b). - Compare these with the Output (a). - Identify missing points, contradictory statements, or factual errors. - Output (b) must address the Question in the same way as the Output (a).
<|im_end|>
<|im_start|>user
I require an assessment of whether Output (b) correctly conveys the core meaning of Output (a). I'll provide you with a question and two model outputs. Your task is to evaluate and return either Output (a) or Output (b), based on the scoring criteria.

Question

```
{
    "question": ""{Question}""
}
```

Model Outputs

Here are the unordered outputs from the models. Each output is associated with a specific model, identified by a unique model identifier.

```
{
    {
        "model_identifier": "m",
        "output": ""{Output (a)}""
    },
    {
        "model_identifier": "M",
        "output": ""{Output (b)}""
    }
}
```

What's your evaluation, Output (a) or Output (b)?

<|im_end|>

B Generated Question-Answer Pair Example from PubMed Publications

QA Example from PubMed

Question:

What role do anti-iNKT TCR antibodies play in activating iNKT cells?

Answer:

Anti-iNKT TCR antibodies can activate iNKT cells by binding to their TCR, which leads to crosslinking and activation. This process can enhance the cytotoxic activity of iNKT cells against target cells, particularly those expressing Fc gamma receptors.

C Expert-Wise Attention

Given a question embedding E_q^i and expert embeddings E_e^j, we define the expert-wise attention mechanism as follows:

$$e_{ij} = \sigma\left(W \cdot \left[E_q^i, E_e^j\right] + b\right) \tag{7}$$

$$\alpha_{ij} = \frac{\exp\left(\sigma\left(W \cdot \left[E_q^i, E_e^j\right] + b\right)\right)}{\sum_{k \in E_e} \exp\left(\sigma\left(W \cdot \left[E_q^i, E_e^k\right] + b\right)\right)} \tag{8}$$

$$Z_i = \sum_{j \in E_e} \alpha_{ij} E_e^j \tag{9}$$

where σ denotes the $ReLU$ activation function, and $[.,.]$ represents embedding concatenation. Furthermore, we concatenate Z_i with each expert embedding E_e^j and pass it through an MLP to obtain the output probability:

$$P\left(E_q^i, E_e^j\right) = \phi\left(\left[Z_i, E_e^j\right]\right) \tag{10}$$