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 cost-sensitive domains such as drug discovery and rare disease research due to the lack of expert knowledge. 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 selectively identifies and queries the most appropriate expert from a team, taking into account each expert's availability, knowledge boundaries, and consultation costs. We train PU-ADKA using simulations on PubMed data and validate it through both controlled expert interactions and real-world deployment with a drug development team, demonstrating its effectiveness in enhancing LLM performance in specialized domains under strict budget constraints. In addition to outlining our methodological innovations and experimental results, we introduce a new benchmark dataset, CKAD, for cost-effective LLM domain knowledge acquisition to foster further research in this challenging area[1].

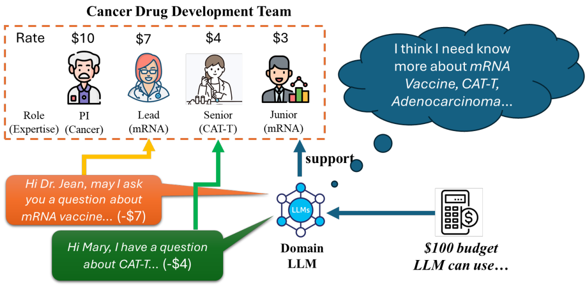

Figure 1: Domain LLM Knowledge Acquisition via Cost-Efficient, Expert-Involved Interaction. The diagram depicts how PU-ADKA selectively engages domain experts with varying expertise and costs to acquire knowledge efficiently within a limited budget.

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; Pal et al., 2024; Yao et al., 2025). However, these gains are not uniformly observed across all domains. In highly specialized and cost-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) (Ouyang et al., 2022; Kaufmann et al., 2023) 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 (Szymanski et al., 2025; Dhar, 2024). 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 Positive Unlabeled Active Domain Knowledge Acquisition (PU-ADKA), which is designed to selectively engage with domain experts and 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

[1]Code and data are included in the submitted supplementary files and will be publicly released after the review process.

released PubMed publications (PubMed, 2024), 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) (Patel et al., 2025). Instead of relying on static, pre-existing datasets, PU-ADKA selectively 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.

Our contributions in this paper are threefold and can be summarized as follows:

- We propose PU-ADKA, a cost-aware framework that strategically selects and queries domain experts by considering their availability, knowledge scope, and consultation cost, in order to enhance LLM performance under limited expert access and fixed budget constraints.

- We introduce the Cost-effective Knowledge Acquisition Dataset (CKAD), a new benchmark for LLM domain knowledge acquisition, to foster further research in the area of domain-specific LLM enhancement.

- We empirically validate the effectiveness of PU-ADKA through both simulation evaluation and a real-world cancer drug development study. The latter experiment involves a drug development team in which five experts with diverse backgrounds participate. The results show that PU-ADKA is promising in enhancing domain LLMs within a fixed budgetary restriction.

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. 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, PU-ADKA, overcomes these shortcomings by redefining expert knowledge acquisition as a budget-constrained optimization task, engaging experts based on their knowledge, cost, and availability, thereby transitioning from static data-driven adaptation to 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; Wang et al., 2024) 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; Li et al., 2022) 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, PU-ADKA integrates active learning with strategic expert collaboration, emphasizing both data sample selection based on the potential to update the model and efficient engagement of experts, balancing cost against their competency and availability.

3 Methodology

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 acquire expert-labeled data for finetuning 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:

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

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.

Table 1: Notations.

Notation	Description
B	Total annotation budget available.
D_{tr}	Unlabeled question pool for training.
D_{te}	Target test set for evaluation.
q_i	The i-th question in the unlabeled pool.
e_j	The j-th domain expert.
f	Allocation function assigning questions to experts.
θ	Base language model.
θ_S	Fine-tuned model using selected question-expert pairs S.
S	Selected set of question–expert pairs for annotation.
$c(q_i, e_j)$	Cost for expert e_j to label question q_i.
x_k^p	Positive question–expert pair used in PU learning.
x_k^u	Unlabeled question–expert pair used in PU learning.
π_p	Prior probability of a positive sample in PU learning.
g	Expert-wise attention network.
$l(\cdot, \cdot)$	Surrogate loss function (e.g., zero-one loss).
Γ_j^t	Number of times expert e_j has been selected up to time t.
w_j^t	Sampling weight of expert e_j at time t.
r_t	Reward at time step t in multi-agent RL.
ϕ_i	Diversity score for question q_i.
$d(E_q^i, E_q^z)$	Distance between question embeddings i and z.
Z_i	Expert-wise representation of question q_i.

3.2 Simulation Environment Construction

To facilitate our study, we introduce a novel benchmark dataset, CKAD, designed to simulate biomedical expert consultations and domain knowledge acquisition for LLMs. This dataset is constructed by strategically leveraging PubMed articles published after the knowledge cutoff date of the base model, ensuring that the selected content represents genuinely novel information. To further isolate new knowledge from prior model capabilities, we implement a temporal knowledge separation mechanism that enforces strict chronological boundaries between the base model's existing knowledge and the newly acquired domain content. This is achieved through three key components detailed below:

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.

Dataset Curation: We construct CKAD from 2024 PubMed Central (PMC) (PubMed, 2024), extracting question-answer (QA) pairs using GPT-4o-2024-08-06 (OpenAI, 2024). For each paper, five mechanism-focused QA pairs are generated using prompting[2] and manually validated. To establish a well-isolated environment for assessing knowledge acquisition, we filter out QA pairs that can be answered by the base model. This process results in a final dataset of *48,219 QA pairs* (the base model cannot correctly answer) representing post-2023 knowledge. To assess the quality of our dataset, we conduct a human evaluation on 100 randomly sampled QA pairs. Two PhD researchers with biomedical backgrounds independently scored each QA pair on a 1–5 scale[3]. The average score is 3.85, and Cohen's Kappa (McHugh, 2012) between the evaluators is 0.73, reflecting high data quality and strong human agreement.

Expert Simulation. To simulate realistic annotation constraints, we construct a binary expert capability matrix $A \in \mathbb{R}^{Q \times N}$, where $A_{ji} = 1$ indicates that expert e_j is assumed to be capable of annotating question q_i, and 0 otherwise. This matrix is used to restrict which expert–question pairs are considered valid during simulation. Without such a constraint, every expert would be able to annotate every question, leading to minimal variation in annotation quality across experts—even for questions unrelated to their domain expertise. To construct A, we use GPT-4o-2024-08-06 to analyze each expert's publications and determine their capacity to annotate specific questions. The top 20 authors ranked by publication count are used as proxy experts. Each expert is assigned a per-

[2]The detail of question-answer extraction prompt is provided in Appendix C.

[3]Quality scoring form is depicted in the Appendix I

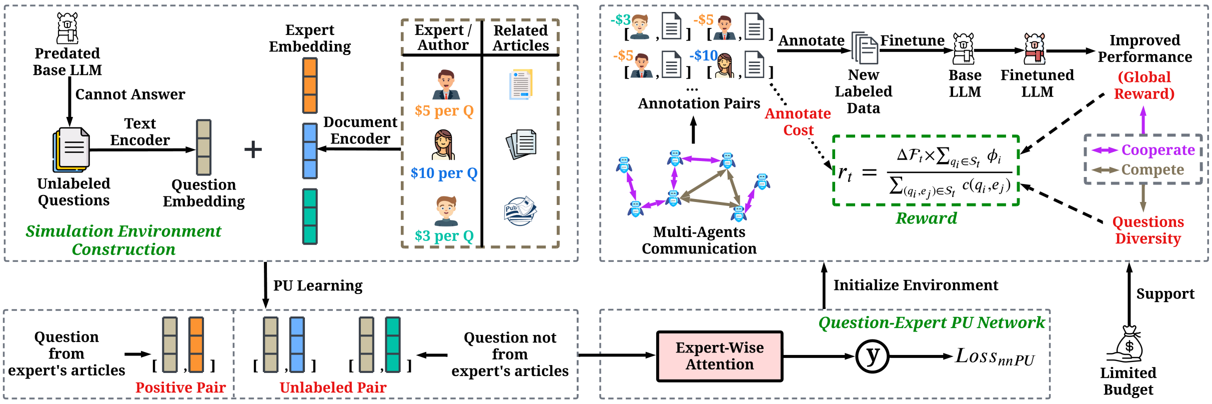

Figure 2: Illustration of our proposed PU-ADKA framework. Given an unlabeled question pool and a team of experts with varying expertise and cost, the question-expert PU learning network identifies the experts that can annotate specific questions based on limited positive examples. A multi-agent reinforcement learning module then selects which questions to annotate and assigns them to appropriate experts under a fixed budget. This process enables efficient acquisition of domain-specific knowledge for the base LLM through expert-in-the-loop supervision.

question annotation rate, determined proportionally by the cumulative impact factor of their publications (Clarivate, 2025).

3.3 Positive Unlabeled Active Domain Knowledge Acquisition

In this section, we present our Positive-Unlabeled Active Domain Knowledge Acquisition (PU-ADKA) framework, which selectively engages domain experts to acquire targeted feedback for improving LLM performance in specialized domains. PU-ADKA comprises two components: (1) Question–Expert Matching, formulated as a Positive-Unlabeled (PU) learning problem to model expert suitability; and (2) Multi-Agent Reinforcement Learning, which selects question–expert pairs under budget constraints. We elaborate on each component below.

3.3.1 Expert Allocation with Positive Unlabeled Learning

Motivation. A key challenge in modeling expert–question suitability lies in the absence of explicit supervision: we can identify which expert authored the source publication from which a question is derived, and thus assume they are qualified to answer it; however, we cannot assume that all other experts are unqualified. This makes standard binary classification infeasible. To address this, we frame the question-expert matching task as a Positive–Unlabeled (PU) learning problem. 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. If q_i does not come from e_j's paper, we do not treat (q_i, e_j) as a negative pair—instead, it remains ***unlabeled***, since the expert may still be qualified. For example, a scholar specializing in cancer NK cells may be able to answer a sepsis-related question involving extracellular vesicles, even without directly publishing in the sepsis domain.

Model Training. 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. To train our PU model to estimate expert knowledge boundary, we employ an expert-wise attention mechanism[4] g 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_p denotes positive class prior ($\pi_p = 0.1$ 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-expert pairs in the labeled positive set and the unlabeled set, respectively.

[4]The attention network is detailed in Appendix D

3.3.2 Domain Knowledge Acquisition via Multi-Agent Reinforcement Learning

The PU learning module is designed to estimate how well each expert aligns with a given question. This section builds on these estimates to select expert–question pairs for annotation under budget constraints, aiming to maximize domain knowledge acquisition.

Motivation. Effective knowledge acquisition requires selecting questions that are not only informative individually, but also complementary as a set. This necessitates modeling dependencies among questions—two high-value questions may become redundant when answered together. For example, questions about *extracellular vesicles* in different disease contexts may overlap in the knowledge they elicit. Single-agent or greedy methods typically overlook such redundancy, leading to inefficient use of limited annotation budgets. To address this, we formulate the question selection as a multi-agent reinforcement learning (RL) problem, where each agent selects a question–expert pair while coordinating through shared rewards. This enables the model to account for inter-question dependencies and optimize the utility of the selected set.

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 weight w_j^t 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.

Multi-Agent 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 $e_1, e_2, \dots, e_h$ 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:

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

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.

Multi-Agent 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 current 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.

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

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

where s_{t+1} denotes the next state, γ is the discount factor, $Q(s, u; \theta)$ is the action-value function, Ω_t and Ω'_t represent 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 the 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 Setup

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.

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. All implementations are conducted with Pytorch(Paszke et al., 2017), PEFT(Mangrulkar et al., 2022) and Transformers(Wolf et al., 2020) on a computation node configured with a 64-core CPU and four 80GB H100 GPUs.

4.2 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 question selection used in baselines:

RAND - 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[5] and quality scorer [6] to score each unlabeled questions.

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. (2024) utilizes gradient similarity to evaluate the potential contribution of each question to the model's performance.

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.

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

Each baseline represents a specific combination of question selection and expert allocation methods, providing a meaningful benchmark against which our proposed approach can be evaluated.

4.3 Evaluation Benchmarks and Metrics.

To ensure a clean evaluation of knowledge acquisition, our CKAD dataset consists of general disease mechanism question-answer pairs[7] that cannot be answered by base LLM (Llama2-7B) initially (i.e., the initial answerable rate is 0). During the simulation training stage, we employ two advanced models, GPT-4o-2024-08-06 (OpenAI, 2024) and GPT-4-Turbo (Achiam et al., 2023), as

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

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

[7] Dataset statistics are provided in Appendix A.

Table 2: Overall performance comparison on CKAD dataset. The best result is highlighted in **bold**, and the second-best is underlined.

Expert Allocation	Question Selection	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

judge models. The evaluation metrics include: **Win Rate** (WR), which measures the percentage of instances where the judge LLM determines that the model-generated answer adequately captures the core meaning of the golden answer; and **Length-Controlled Win Rate** (LC_WR) (Dubois et al., 2024a), a variant of WR that filters out samples with large answer length discrepancies between the model output and the golden answer, helping to control for verbosity bias during evaluation.

Additionally, following (Wang et al., 2023), we conduct human-involved experiments to validate the effectiveness of our method. The expert team consists of three sepsis specialists and two cancer specialists, representing different levels of expertise. Among them, one expert is a medical doctor, and the remaining four are PhD students[8].

4.4 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 WR of 18.2% and an LC_WR of 25.65%. When evaluated by the GPT-4-Turbo model, it records a WR 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 WR, 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 WR at 12.1% and 10% in both settings. Furthermore, the minimal baseline performance under fully random conditions, with WR 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)

4.5 Human Involved Validation

To further substantiate the robustness of our method, PU-ADKA, we implement it within a pro-

[8]Detailed information about the human experts is provided in Ethics Statement.

fessional biomedical team of experts under a simulated budget constraint of $100 per game. The cost of human experts is 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 labeled question. We assess the performance in terms of WR and LC_WR 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.

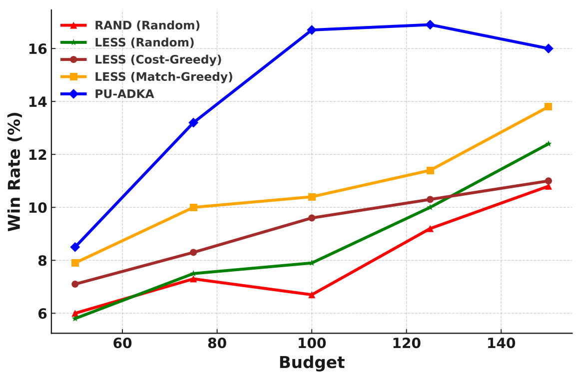

Figure 3: Performance comparison under different budgets ($) evaluated by GPT-4-Turbo

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

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

The results reveal that PU-ADKA notably surpasses the most competitive baseline, LESS (Match-Greedy), by margins of 2.7% and 3.1% in WR and LC_WR, 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.

4.6 Ablation Study

4.6.1 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 perform a series of ablation studies. These studies are conducted on the QA dataset, with GPT-4-Turbo serving as the judge. We explore 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.

4.6.2 Performance under Different Budgets

Following (Hacohen et al., 2022; Li et al., 2022), we evaluate 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. This observation 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

We propose PU-ADKA, a cost-aware active learning framework that enhances LLMs by selectively engaging domain experts based on their expertise, availability, and annotation cost. PU-ADKA improves budget efficiency and LLM performance, as validated through simulations and real-world biomedical tasks. The release of the CKAD dataset supports further research on domain-specific LLM tuning. In future work, we plan to explore alternative backbone models and extend PU-ADKA to specialized domains beyond biomedical domain.

Limitations

• *Scalability with Increasing Data and Experts.* As the number of unlabeled data points and available 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 do not extensively experiment with varying agent numbers due to the high computational cost associated with training and coordination. Additionally, we do 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.

• *Backbone Diversity and Model Size.* We adopt LLaMA2-7B as the fixed backbone to ensure consistent evaluation. However, this limits the exploration of PU-ADKA's effectiveness across other model families and sizes. Future work should investigate its adaptability to diverse architectures, including larger models, instruction-tuned variants, and domain-specific LLMs.

Ethics Statement

All human annotation work in this study is conducted by domain experts in an external biomedical research group. Importantly, none of the experts are listed as co-authors of this paper. The process is coordinated by a biomedical Principal Investigator (PI), who is a co-author, but does not participate in any annotation work directly. The experts include one medical doctor and four PhD-level biomedical researchers, and they are blind to the study's hypotheses, model design, and experiments.

All annotation work is performed during the experts' regular paid working hours as part of their institutional responsibilities, and no additional compensation is provided. To simulate annotation cost in our experiments, we adopt a relative cost scheme based on typical salary ratios across seniority levels (e.g., doctor : senior PhD : junior PhD = 5 : 2 : 1), without disclosing any actual salary details. All experts agree that their annotations will be used in this study and released alongside the dataset. The content they annotate is derived entirely from publicly available biomedical literature and contains no personal or sensitive information. Accordingly, no additional ethics board review is required under the ACL Ethics Policy.

References

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

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.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

UU Dhar. 2024. Enabling domain expert evaluation of emerging ai technologies in healthcare settings.

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.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024a. Length-controlled alpacaeval: A simple way to debias automatic evaluators,. *URL https://arxiv. org/abs/2404.04475*.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024b. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

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*.

Guy Hacohen, Avihu Dekel, and Daphna Weinshall. 2022. Active learning on a budget: Opposite strategies suit high and low budgets. *arXiv preprint arXiv:2202.02794*.

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.

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*.

Shibo Li, Jeff M Phillips, Xin Yu, Robert Kirby, and Shandian Zhe. 2022. Batch multi-fidelity active learning with budget constraints. *Advances in Neural Information Processing Systems*, 35:995–1007.

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*.

Sanket Mangrulkar, Kaustubh Somasundaram, and Akhilesh Shrivastava. 2022. Peft: Parameter-efficient fine-tuning. *Hugging Face*. https://github.com/huggingface/peft.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

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.

Soumen Pal, Manojit Bhattacharya, Sang-Soo Lee, and Chiranjib Chakraborty. 2024. A domain-specific next-generation large language model (llm) or chatgpt is required for biomedical engineering and research. *Annals of biomedical engineering*, 52(3):451–454.

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.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2017. Automatic differentiation in pytorch. In *Advances in Neural Information Processing Systems*.

Sawan Patel, Fred Zhangzhi Peng, Keith Fraser, Adam D Friedman, Pranam Chatterjee, and Sherwood Yao. 2025. Evoflow-rna: Generating and representing non-coding rna with a language model. *bioRxiv*, pages 2025–02.

PubMed. 2024. PubMed Data: Download and Use. https://pubmed.ncbi.nlm.nih.gov/download. Last updated Dec 14, 2024.

Annalisa Szymanski, Noah Ziems, Heather A Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A Metoyer. 2025. Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 952–966.

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*.

Hongru Wang, Rui Wang, Fei Mi, Zezhong Wang, Ruifeng Xu, and Kam-Fai Wong. 2023. Chain-of-thought prompting for responding to in-depth dialogue questions with llm. *arXiv preprint arXiv:2305.11792*.

Jian Wang, Zhe Zhang, and Guosheng Zhao. 2024. Tag-based self-learning task recommendation for mobile crowdsensing via collaborative multi-expert system. *Computer Communications*, 214:260–269.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yang Wu, Huayi Zhang, Yizheng Jiao, Lin Ma, Xiaozhong Liu, Jinhong Yu, Dongyu Zhang, Dezhi Yu, and Wei Xu. 2024. 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*.

Rujing Yao, Yang Wu, Chenghao Wang, Jingwei Xiong, Fang Wang, and Xiaozhong Liu. 2025. Elevating legal llm responses: Harnessing trainable logical structures and semantic knowledge with legal reasoning. *arXiv preprint arXiv:2502.07912*.

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 Dataset Statistics

In this section, we present key statistics of the CKAD dataset in Table 5, which includes question–answer pairs generated from PubMed 2024 articles focused on Sepsis and Cancer NK cell mechanisms.

Table 5: Statistics of CKAD dataset.

Disease Type	Cancer_NK and Sepsis
#Train	38,575
#Dev	4,722
#Test	4,722
#Avg. Tokens in Question	12
#Avg. Tokens in Answer	29

B Generated Question-Answer Examples from PubMed Publications

Cancer QA Example from PubMed

Question:

What role does MHC-I play in modulating NK cell activity against tumor cells?

Answer:

MHC-I molecules on tumor cells can engage with inhibitory receptors on natural killer (NK) cells, such as KIRs and NKG2A, reducing NK cell activation and cytotoxic activity against the tumor.

Sepsis QA Example from PubMed

Question:

How does anti-thrombin administration influence the risk of bleeding complications in sepsis patients?

Answer:

Anti-thrombin administration can increase the risk of bleeding complications by enhancing anticoagulant activity, which may impair the body's ability to form necessary clots and maintain hemostasis.

C Prompts

In this section, we present the detailed prompts used for generating Question-Answer data and the specific prompts employed for model evaluation across all experiments. For evaluation, we slightly rephrase the final prompt instruction—asking the model to choose between output a or b—to simplify post-processing and ensure compatibility with the AlpacaEval (Dubois et al., 2024b) evaluation package. This modification does not affect the evaluation result, which remains to judge whether output b correctly reflects the meaning of output a. Notably, this evaluation does not simply reflect the judge model's general preference between two answers but specifically assesses whether the target model's answer adequately captures the core meaning of the golden answer.

C.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|>

C.2 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|>

D 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:

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

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

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

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

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

E Overall Comparison of Question Selection Methods

Table 6 summarizes the average performance of different question selection methods across all expert allocation strategies. PU-ADKA shows superior results on both WR and LC_WR metrics, with LESS ranking second in overall effectiveness. We therefore use LESS as the default question selection method for baselines in our ablation studies.

Table 6: Average performance of different question selection methods across three expert allocation strategies.

Method	WR.Avg (%)	LC_WR.Avg (%)	Overall.Avg (%)
RAND	6.48	20.40	13.44
DEITA	<u>10.43</u>	21.40	15.92
CHERRY	9.08	21.95	15.52
NUGGETS	9.80	21.17	15.49
LESS	10.00	<u>22.55</u>	<u>16.27</u>
ROSE	9.37	21.95	15.66
PU-ADKA	**17.45**	**26.05**	**21.75**

F Comparison with Fully Annotated Upper Bound

To understand the upper performance bound achievable without cost constraints, we compare PU-ADKA against a setting where all training data are fully annotated (FULL). As expected, FULL yields the highest scores in Table 7, serving as an empirical upper bound. However, PU-ADKA approaches

Table 7: Performance comparison under fixed-budget and fully annotated settings. The FULL setting denotes that all available data are exhaustively annotated without any budget constraints, serving as an empirical upper bound.

Expert Allocation	Question Selection	GPT-4o-2024-08-06	GPT-4-Turbo	GPT-4o-2024-08-06	GPT-4-Turbo	Avg.Length
		WR (%)	WR (%)	LC_WR (%)	LC_WR (%)	-
-	FULL	**22.1** (0.7)	**19.3** (0.9)	**27.8** (1.0)	**28.1** (0.8)	1752
Random	LESS	7.9 (0.2)	7.9 (0.2)	22.0 (1.0)	24.0 (1.1)	2212
Cost-Greedy		12.1 (0.4)	9.6 (0.4)	22.1 (0.8)	21.2 (1.0)	2218
Match-Greedy		12.1 (0.4)	10.4 (0.2)	23.5 (1.0)	22.5 (1.0)	2252
Ours	PU-ADKA	18.2 (0.6)	16.7 (0.4)	25.6 (1.0)	26.5 (0.9)	1781

Table 8: Performance comparison between different encoders used for question representation within PU-ADKA framework.

Encoder	GPT-4o-2024-08-06	GPT-4-Turbo	GPT-4o-2024-08-06	GPT-4-Turbo	Avg.Length
	WR (%)	WR (%)	LC_WR (%)	LC_WR (%)	-
BERT-base	16.3 (0.9)	12.9 (0.7)	24.0 (1.0)	25.4 (1.2)	1967
Llama2-7B (ours)	**18.2** (0.6)	**16.7** (0.4)	**25.6** (1.0)	**26.5** (0.9)	1781

Table 9: Number of annotated QA pairs and evaluation results under different expert allocation strategies (questions selected by LESS; judged by GPT-4o-2024-08-06).

Method	Annotated QA Pairs	WR (%)	LC-WR (%)
Random	312	7.9	22.0
Cost-Greedy	1000	12.1	22.1
Match-Greedy	508	12.1	23.5
Ours (PU-ADKA)	632	**18.2**	**25.6**

Table 10: Data Quality Scoring Form

Score	Description
1	Incorrect or irrelevant.
2	Partially correct, key issues.
3	Correct and main point covered.
4	Correct with minor omissions.
5	Fully correct and complete.

this upper limit while operating under a strict $100 budget—substantially outperforming all baselines in both WR and LC_WR scores. This highlights the effectiveness of PU-ADKA in achieving strong domain adaptation with fewer annotations.

G Effect of Encoder Architecture on Performance

To assess the impact of the encoder architecture on our framework, we compare a standard BERT-base (Devlin et al., 2019) model with our default Llama2-7B encoder. Results in Table 8 show that Llama2-7B consistently outperforms BERT-base encoder across all evaluation metrics.

H Annotation Quantity under Budget Constraints

Table 9 shows that although the Cost-Greedy strategy yields the largest number of annotated QA pairs, it does not achieve the best performance. In contrast, PU-ADKA achieves a balance between annotation quantity and quality: it produces more annotations than Random and Match-Greedy strategies and achieves the highest WR and LC_WR scores. This demonstrates PU-ADKA's effectiveness in utilizing limited budgets to acquire high-quality supervision.

I Data Quality Scoring Form

The quality of each QA pair is scored on a 1–5 scale based on the correctness and completeness of its biomedical mechanistic explanation. The scoring rubric is shown in Table 10.

J Discussion

We adopt the base Llama2-7B model rather than an instruction-tuned variant to ensure a controlled setting where observed improvements are attributable solely to our expert-interaction framework, without influence from pretrained instruction-following capabilities. Besides, GPT-4o-2024-08-06 is used to generate QA pairs based on the factual content of 2024 PubMed articles. This generation process is independent of any model comparison or evaluation. To mitigate potential bias from using a single evaluation model, we also include GPT-4-Turbo as a second judge model.