Cost-efficient Knowledge-based Question Answering with Large Language Models

Junnan Dong[1], **Qinggang Zhang**[1], **Chuang Zhou**[1], **Hao Chen**[1†], **Daochen Zha**[2], **Xiao Huang**[1]

[1] The Hong Kong Polytechnic University
[2] Rice University
{hanson.dong, qinggangg.zhang, chuang-qqzj.zhou}@connect.polyu.hk,
sundaychenhao@gmail.com,daochen.zha@rice.edu, xiaohuang@polyu.edu.hk

Abstract

Knowledge-based question answering (KBQA) is widely used in many scenarios that necessitate domain knowledge. Large language models (LLMs) bring opportunities to KBQA, while their costs are significantly higher and absence of domain-specific knowledge during pre-training. We are motivated to combine LLMs and prior small models on knowledge graphs (KGMs) for both inferential accuracy and cost saving. However, it remains challenging since accuracy and cost are not readily combined in the optimization as two distinct metrics. It is also laborious for model selection since different models excel in diverse knowledge. To this end, we propose `Coke`, a novel cost-efficient strategy for KBQA with LLMs, modeled as a tailored multi-armed bandit problem to minimize calls to LLMs within limited budgets. We first formulate the *accuracy expectation* with a cluster-level Thompson Sampling for either KGMs or LLMs. A context-aware policy is optimized to further distinguish the expert model subject to the question semantics. The overall decision is bounded by the *cost regret* according to historical expenditure on failures. Extensive experiments showcase the superior performance of `Coke`, which moves the Pareto frontier with up to 20.89% saving of GPT-4 fees while achieving a 2.74% higher accuracy on the benchmark datasets.

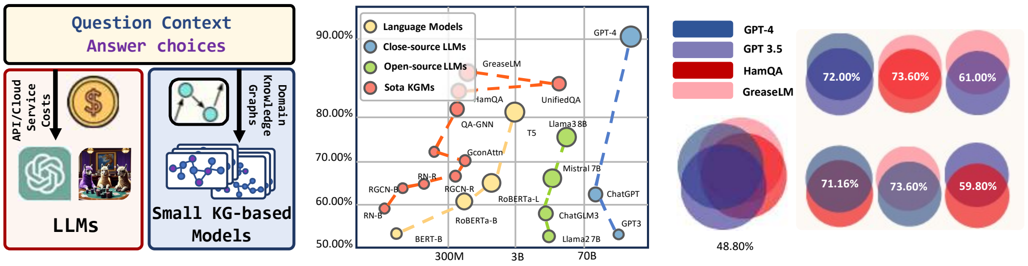
1 Introduction

Knowledge-based question answering (KBQA) has gained significant attention across various specialized domains, e.g., education and medicine [24, 13, 21, 20]. Given a question, the model is required to make inferences based on necessary reasoning background [14]. Inspired by the effectiveness of knowledge graphs (KGs), e.g., ConceptNet [33, 8], where real-world entities are represented in the structural form as (*head entity*, *relation*, *tail entity*) [7], KG-based models (KGMs) have been proposed to leverage KGs for reasoning. Based on the hypothesis that answers could be located multiple hops away from the question concepts [24] in KGs, former studies are mainly dedicated to tracing the trajectory from questions to answers to model the structural information [16, 25, 5, 4], or utilizing graph neural networks (GNNs) to learn the question-specific subgraph from KGs [14, 37, 39].

With the emergence of large language models (LLMs), e.g., ChatGPT and GPT-4 [29], They have shown remarkable performance benefited from the injected knowledge during pre-training [38, 12]. However, it is challenging to adopt LLMs in practice. First, either calling the API or deploying the open-source LLMs with cloud service is prohibitive [2, 9]. GPT-4 is estimated to cost at least thousands of dollars for pilot-scale customer service [6] while Llama3 70B requires unaffordable

† Corresponding Author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).



(a) (b) (c)

Figure 1: A sketched overview of LLMs and small KGMs in (a) We visualize the Acc./Param size of both pipelines of models in (b) The overlaps among different model predictions are shown in (c).

computation resources for small business, e.g., 40G graphics memory, let alone the high throughput scenarios like online e-commerce [30]. Second, LLMs may struggle to identify the correct answer for certain questions due to the lack of particular knowledge that is not covered in their pre-training corpus [1, 32, 15]. They are considered unreliable to assist pedagogical or medical purposes since they could generate hallucination [3, 18] and misleading responses [17, 19].

We illustrate the sketched overview of LLMs and KGMs methods in Figure 1 (a) where LLMs are used in a zero-shot setting with direct prompts, e.g., [`Question:Choices`], and KGMs require the external domain KG as reasoning background. In Figure 1 (b), we visualize the accuracy/parameter size of four types of models. In general, KGMs are far more lightweight considering model size but tend to underperform LLMs overall. Conversely, LLMs are more computationally expensive and may struggle with specific questions that demand knowledge not covered in the pre-training corpus. We are thereby motivated to combine the strengths of LLMs and KGMs through question-specific model selection to improve the Pareto frontier, achieving higher inferential accuracy and lower costs.

However, it is nontrivial for two major challenges. First, inferential accuracy and cost saving are two distinct metrics. It is hard to combine them in the optimization simultaneously. Since more parameters will lead to higher costs, it indicates the importance of a careful selection that balances both exploration and exploitation. Second, selecting the most suitable model for particular questions is laborious. In Figure 1 (c), we showcase the overlaps among several representative models on the benchmark OpenBookQA [28] dataset. While different models focus on various considerations, they may consequently excel in diverse knowledge and question types. For example, HamQA [14] focuses on hypernymy in the questions, i.e., cats and mammals. This makes it expensive to incorporate expertise according to the specialty of models to answer diverse real-world domain-specific questions.

To this end, we present a novel cost-efficient strategy to leverage LLMs for KBQA, i.e., `Coke`. It is modeled as a tailored multi-armed bandit (MAB) problem, which is trained to assign the most promising model for each question within a considerably limited budget. Specifically, we assemble two sets of base models, i.e., LLMs and KGMs to balance both inferential accuracy and cost saving. (i) we first model the preliminary selection with a cluster-level Thompson Sampling. It suggests the *accuracy expectation* of choosing either LLMs or KGMs based on historical success and failure at Beta distribution. (ii) A context-aware policy is learned to further distinguish the most suitable model. This could effectively assign the corresponding expert for the given question semantics. (iii) The overall decision making is bounded by the *cost regret*. It indicatively constrains the selection based on the cumulative expenses incurred from failures of the current model.

Contributions.

- ▶ We formally define the task of optimizing both inferential accuracy and cost for KBQA with LLMs.
- ▶ A novel cost-efficient strategy, `Coke`, is presented to automatically assign the most promising model for particular questions. It could effectively make reliable decisions considering both *inferential accuracy* and *cost saving*, making a balance of *exploration* and *exploitation* during selections.
- ▶ Extensive experiments over three domain-specific benchmark datasets demonstrate the superiority of our proposed framework, moving the Pareto frontier to achieve higher accuracy and lower costs.

2 Problem Formulation

2.1 Task Definition

We adopt lowercase alphabets for scalars (e.g., m), boldface lowercase letters for vectors(e.g., $\mathbf{q}$) and boldface uppercase ones for matrices (e.g., $\mathbf{Q}$). The decorated letters are used for sets, e.g., $\mathcal{A}$. We assemble two clusters of models, i.e., $\mathcal{C} = \{c_L, c_K\}$ for sets of LLMs and KGMs, respectively. All the candidate models in $\mathcal{C}$ are denoted as $\mathcal{M} = \{m_1, m_2..m_n\}$. The knowledge graph used in KGMs is expressed as $\mathcal{G}$ where real-world entities e are represented in the form of triples as (e_h, r, e_t).

Given domain-specific questions $Q = \{q_1, q_2..q_u\}$, and several candidate models $\mathcal{M}$, we aim to optimize the framework to identify the most promising model within a limited budget $\mathcal{B}$ to invoke LLMs. The overall performance is evaluated by both inferential accuracy and cost saving. We aim to maximize the accuracy comparing prediction $\hat{y}_i$ and ground truth y_i, i.e., $max(f_{acc}(\hat{y}_i|y_i))$, and minimize the costs, i.e., $min(f_{cost}(p(m)|q_i))$ where $p(m)$ is the unit cost of model m.

2.2 Performance Evaluation

For fair comparisons, we divide the overall evaluation of Coke against state-of-the-art baselines into two parts considering both *inference accuracy* and *cost saving*.

Inferential accuracy: The performance of KBQA task itself is evaluated by the overall accuracy of the model prediction compared with ground truths, i.e., {0, 1} for each question indicates wrong/correct. The accuracy is expected to be as high as possible to correctly answer more questions.

Cost Saving: The cost of using particular models is often case by case. It could involve many aspects, e.g., power consumption, GPU and graphics memory costs, cloud service charges and token-level API expenses, etc. In this paper, we instantiate this metric for Coke in two ways; one can define the cost in other ways under our framework. (i) *API fees* ($). This intuitively indicates the cost of money particularly for comparing with a series of LLMs from OpenAI, e.g., GPT3.5 and GPT4. (ii) *Calls* (times). We generalize the evaluation to open-source LLMs like Llama and ChatGLM. It indicates the number of times that we invoke the LLMs. Since KGMs are considered far more cheaper for local and cloud implementation, the metric of calls is also expected to be as few as possible.

3 Approach: Coke

To achieve an effective model selection, we mainly aim to answer two research questions: (i) ***How can we balance the exploration and exploitation to find the best model for both accuracy and cost saving?*** Real-world questions are difficult to manually identify to apply LLMs or KGMs. We usually value the prior knowledge to select the most accurate and inexpensive model at present greedily, i.e., *exploitation*. However, we also wish to obtain sufficient posterior experiences from under-explored models so far, i.e., *exploration* to find more models with a superior performance-to-price ratio. (ii) ***How can we automatically distinguish the most promising expert models for different questions?*** The particular types of required knowledge vary among questions. This suggests a careful selection to leverage the specialized expertise from different models to make correct inferences. To this end, we design a novel cost-efficient framework, Coke, to automatically select question-specif models.

We formulate the automatic model selection as a multi-armed bandit (MAB) problem. Specifically, in each iteration $k \leq \mathcal{K}$, our policy is presented with the choice of selecting one model $m \in \mathcal{M}$ out of $\mathcal{N}$ candidate models, referred to as $\mathcal{N}$ *arms*. Let $r_m^k \in \{0, 1\}$ denote the reward obtained by selecting the model m at the iteration k, where $m \in \mathcal{M}$. For each given question q, if the chosen model m can correctly answer it, the policy will receive a positive reward of 1, and 0 vice versa. We aim to maximize the cumulative rewards in $\mathcal{K}$ iterations and formulate the objective as follows:

$$\max \sum_{k=1}^{\mathcal{K}} r_m^k. \tag{1}$$

In each iteration, the selection is inherently associated with an expectation function $\mathbb{E}(\cdot)$ to capture the implicit linear correlation between the given question q and the potential reward r, indicating the likelihood of success by choosing one particular model m.

3.1 Accuracy-Cost Trade-off for KBQA with LLMs

To answer the two aforementioned questions, we decompose the expectation $\mathbb{E}(r_m^k|q)$ into three aspects. First, we calculate the *accuracy expectation* of one cluster, i.e., LLMs or KGMs, for better inferential performance(*). Second, we design a context-aware expert distinguishing to present the question embedding to the policy and find the best arm with implicit expertise (**), which further increases the probability of achieving higher accuracy for different question contexts. Finally, we constrain the decision-making with a *cost regret*. This is introduced to punish the model which wastes more on failures (***). The overall expectation is correspondingly formulated as follows.

$$\mathbb{E}(r_m^k|q_k) = \underbrace{\mathbb{E}_c(r_c|\theta_c)}_{(*)} + \underbrace{\mathbb{E}_a(r_a|q_k)}_{(**)} - \lambda \cdot \underbrace{\mathcal{R}_a(\mathcal{B}, p(a), q_i)}_{(***)}, \tag{2}$$

where $r_m^k \in \{0, 1\}$, r_c and r_a are the decomposed rewards for cluster sampling in terms of accuracy and arm selection considering knowledge expertise for particular questions. $\mathcal{B}$ indicates the limited budget allocated to invoke LLMs. Details are described hereunder.

3.2 Accuracy-Encouraged Cluster Expectation (*)

To encourage the policy to select more promising models, we first establish a higher-level expectation concerning the overall accuracy performance of KGMs and LLMs. Inspired by the traditional Thompson Sampling, which iteratively samples the best arm based on historical information, i.e., success and failure, we thereby design a tailored cluster-level Thompson Sampling to evaluate the expectation of choosing one particular cluster c. It presents a dynamic approach where the selections evolve over iterations based on success and failures, gradually converging towards optimal selections as more questions are encountered. This could also effectively embody the *exploration-exploitation* trade-off inherent in our model selection scenarios for the sake of accuracy.

Specifically, it is established based on the prior knowledge as *(success, failure)* of each cluster, instantiated by a Beta distribution of $Beta(\alpha, \beta)$. We define this pair of conjugate prior below.

Definition: conjugate prior of α and β. *Given a cluster $c \in \{c_L, c_K\}$, let α_c and β_c denote the success and failure prior respectively for c. The pair of (α_c, β_c) forms a conjugate prior. It facilitates the efficient update of fundamental beliefs about the performance of each cluster c during selections.*

In general, we consider our cluster-level accuracy expectation $\mathbb{E}_c$ as a likelihood of success for each cluster by randomly sampling an indicator θ_c^k in iteration k from the distribution of $Beta(\alpha_c^{k-1}, \beta_c^{k-1})$ based on the observation of success and failure in the former $(k-1)$ rounds.

$$\mathbb{E}_c(r_c^k|\theta_c) \propto \theta_c \sim (\alpha_c^{k-1}, \beta_c^{k-1}), \tag{3}$$

where θ_c is distributed approximately uniformly across the entire interval, resulting in a uniform *exploration* when a particular cluster has not been extensively sampled. Otherwise, if cluster c has been sufficiently selected and the performance turns out to be satisfying with more success times, i.e., larger α_c, the corresponding expectation associated with θ_c will be more likely to be higher to facilitate a reliable *exploitation*. When $k = 0$, we value the prior knowledge of cluster expectation before any observations have been made. In our paper, we utilize the average reported performance of each arm within the cluster as the prior for θ_c. This ensures an empirically grounded initial belief about the success probability of the cluster, as well as for subsequent Bayesian inference.

$$prior(\theta_c^k) \sim \frac{\mathbf{\Gamma}(\alpha_c^{k-1} + \beta_c^{k-1})}{\mathbf{\Gamma}(\alpha_c^{k-1}) \cdot \mathbf{\Gamma}(\beta_c^{k-1})} \cdot \theta_c^{\alpha_c - 1} \cdot (1 - \theta_c)^{\beta_c^{k-1} - 1}, \tag{4}$$

where $\Gamma(\cdot)$ is the Gamma function. Based on this, we consider the cluster with the largest θ_c^k in iteration k as the best cluster c^* for current question q, where the arm models within this particular cluster will have higher chances of answering this question. A reward $r_c^k \in \{1, 0\}$ will then be given if c^* can/cannot make the correct prediction. Correspondingly, the posterior distributions for all the historically selected clusters will be updated when $0 < k \leq \mathcal{K}$. The history of cluster sampling and the posterior updating is formulated as follows, respectively.

$$\mathcal{H}_c^{k-1} = \{c_n^\tau, \alpha_c^{k-1}, \beta_c^{k-1}, \mathbf{r}_c^{k-1}, \tau = 1, 2...k-1, n = 1, 2...N\}$$

$$posterior(\theta_c^k) \sim \frac{\mathbf{\Gamma}(\alpha_c^{k-1} + \beta_c^{k-1} + 1)}{\mathbf{\Gamma}(\alpha_c^{k-1} + r_c^{k-1}) \cdot \mathbf{\Gamma}(\beta_c^{k-1} + 1 - r_c^{k-1})} \cdot (\theta_c^k)^{\alpha_c^{k-1} + r_c^{k-1} - 1} \cdot (1 - \theta_c^k)^{\beta_c^{k-1} - r_c^{k-1}}. \tag{5}$$

While the exact sequence of cluster selections may vary between runs, the overall behavior will eventually converge to optimal cluster selections over time in our modeled MAB problem, especially as more questions are presented and more historical success/failure information is observed.

Definition: Selection Regret. *In iteration k, we denote the real-selected arm as a_k, the annotated best arm as a_k^*, which can answer the question with the lowest costs. We refer to the expectation differences between a_k and a_k^* as the selection regret for current iteration k.*

Specifically, we give the overall selection regret bounds for the cluster-level Thompson Sampling as follows. Given the selected arm a_k and the historical information H_k up to iteration k, $\mathbb{E}_c$ is bounded as follows: (The detailed proof of confidence bound is provided in the **Appendix** Section A)

$$SR(\mathcal{K}) \leq 2\gamma + 2\sum_{k=1}^{\mathcal{K}} \mathbb{E}[r(a_k, H_{k-1})]. \tag{6}$$

In general, the bounds are obtained from the following derivations. Initially, we establish the upper and lower confidence bounds as explicit functions of the arm a_k and history H_{k-1} respectively, denoted as $U(a_k, H_{k-1})$ and $L(a_k, H_{k-1})$. For some $\gamma > 0$ and the number of arms $\mathcal{N}$, the specific form of functions U and L is irrelevant as long as they satisfy the following properties:

$$\begin{aligned}
&\forall a,\ k, \quad \mathbb{E}\big[\big[U(a, H_{k-1}) - \mu(a)\big]^-\big] \leq \frac{\gamma}{\mathcal{K} \cdot \mathcal{N}}, \\
&\forall a,\ k, \quad \mathbb{E}\big[\big[\mu(a) - L(a, H_{k-1})\big]^-\big] \leq \frac{\gamma}{\mathcal{K} \cdot \mathcal{N}}.
\end{aligned} \tag{7}$$

3.3 Context-Aware Expert Distinguishing (**)

To answer the second question, we further deepen our MAB problem as a contextual variant. In this part, the expectation is highly related to the question semantics. We aim to automatically learn from the vector representation of questions, e.g., $\mathbf{q} \in \mathbb{R}^d$, d for dimension, and effectively identify the corresponding expert model to answer it. The embedding is uniformly obtained by applying a lightweight pre-trained language model RoberTa [27]. To this end, we design the expectation function $\mathbb{E}_a(r_a^k|q^k)$ to effectively capture the linear correlation between $\mathbf{q}$ and r_a in iteration k.

$$\mathbb{E}_a(r_a^k|q^k) = \mathbf{q}^k \times \boldsymbol{\mu}_a^{k-1} + \eta_a^{k-1}, \tag{8}$$

where $\boldsymbol{\mu}_a^{k-1} \in \mathbb{R}^{1\times d}$ is a learned vectored parameter associated with each arm a in $k-1$ steps. We introduce η_a^{k-1} as a trainable noise at Gaussian distribution $\hat{\mathcal{N}}(0, (\Delta^{(n)})^2)$ to balance the *exploration* and *exploitation*. We maximize $\mathbf{q}^k \times \boldsymbol{\mu}_a^{k-1}$ to encourage the exploitation [10]. The tight correlations are established among the given question, the history information, i.e., $\mathcal{H}_a^{k-1}$, including all the questions $\mathbf{Q}_a^{k-1} \in \mathbb{R}^{(k-1)\times d}$ answered by arm a and the rewards received $\mathbf{r}_a^{k-1} \in \mathbb{R}^{1\times(k-1)}$ in k-1 iterations. We could observe the history $\mathcal{H}$ for each arm for abundant information for reference as:

$$\mathcal{H}_a^{k-1} = \{a_n^\tau, \mathbf{Q}_a^{k-1}, \mathbf{r}_a^{k-1}, \tau = 1, 2...k-1, n = 1, 2...N\} \tag{9}$$

Specifically, we update $\boldsymbol{\mu}_a^k$ based on the $\mathcal{H}_a^{k-1}$ with a typical ridge regression $f(\cdot)$.

$$\begin{aligned}
f(\mathbf{Q}_a^{k-1}, \mathbf{r}_a^{k-1}) &= \sum_{k=1}^{K} (\mathbf{r}_a^{k-1} - \mathbf{Q}_a^{k-1}\boldsymbol{\mu}_a^k)^2 + \sigma \parallel \boldsymbol{\mu}_a^k \parallel_2^2 \\
\rightarrow f(\boldsymbol{\mu}_a^k) &= (\mathbf{r}_a^{k-1} - \mathbf{Q}_a^{k-1}(\boldsymbol{\mu}_a^k)^\top(\mathbf{r}_a^{k-1} - \mathbf{Q}_a^{k-1}\boldsymbol{\mu}_a^k) + \sigma^b(\boldsymbol{\mu}_a^k)^\top \boldsymbol{\mu}_a^k,
\end{aligned} \tag{10}$$

where we introduce the L2 normalization to ensure the reversibility of $\mathbf{Q}_a^{k-1}$ in addition to the original ordinary least square loss. σ solves the over-fitting problem by adopting a suitable λ. To find the optimal value of $\boldsymbol{\mu}_a^k$ that minimizes the cost function, we differentiate the formula with respect to $\boldsymbol{\mu}_a^k$, set the derivative equal to zero, and solve it. This yields the following update equation:

$$\boldsymbol{\mu}_a^k = \left((\mathbf{Q}_a^{k-1})^\top \mathbf{Q}_a^{k-1} + \sigma_a \mathbf{I}\right)^{-1} (\mathbf{Q}_a^{k-1})^\top \mathbf{r}_a^{k-1}, \tag{11}$$

where $\mathbf{I} \in \mathbb{R}^{d\times d}$ is an identity matrix. To facilitate exploration on less explored arms, we adopt an upper confidence bound for exploration-exploitation trade-off by introducing η_a^{k-1}. For any $\delta > 0$ with the probability at least $(1-\delta)$, the expectation $\mathbb{E}_a(r_a^k|\mathbf{q}^k)$ is bounded by a confidence interval:

$$\mathbf{q}^k \boldsymbol{\mu} a^{k-1} - \gamma \times f\gamma(\mathbf{Q}_a^{k-1}) \leq \mathbb{E}_a \leq \mathbf{q}^k \boldsymbol{\mu} a^{k-1} + \gamma \times f\gamma(\mathbf{Q}_a^{k-1}), \tag{12}$$

where γ is a constant value, i.e., $\gamma = 1 + \sqrt{ln(2/\delta)/2}$. Also, we can derive the correlation term $f\gamma(\mathbf{Q}a^{k-1})$ as $f\gamma(\mathbf{Q}_a^{k-1}) \triangleq \sqrt{\mathbf{q}^k \left((\mathbf{Q}_a^{k-1})^\top \mathbf{Q}_a^{k-1} + \sigma\mathbf{I}\right)^{-1} (\mathbf{q}^k)^\top}$. Hence, through learning from the current question $\mathbf{q}^k \in \mathbb{R}^d$ and all historically assigned questions $\mathbf{Q}_a^{k-1} \in \mathbb{R}^{(k-1)\times d}$ for arm a, we can simply derive the equation to appropriately update the noise term η_a^{k-1} for next iteration.

$$\eta_a^k = \quad \gamma \times \sqrt{\mathbf{q}^k \left((\mathbf{Q}_a^{k-1})^\top \mathbf{Q}_a^{k-1} + \sigma_a\mathbf{I}\right)^{-1} (\mathbf{q}^k)^\top}. \tag{13}$$

When an arm a with larger $\mathbf{q}^k \boldsymbol{\mu}_a^{k-1}$ is selected, this reflects an *exploitation* process. Similarly, when the model chooses an arm with larger η_a^{k-1} learned in previous iterations, this variance shows an *exploration* process since the model performs few selections of the current arm. Thus, jointly maximizing $(\mathbf{q}^k \times \boldsymbol{\mu}_a^{k-1} + \eta_a^{k-1})$ helps us for more promising expert models subject to question q^k.

In conclusion, guided by the objective of maximizing cumulative rewards r_m^k, we concentrate on the sub-target of finding contextually best arm as the expert to correctly answer the question by prioritizing the arm with higher expectation $\mathbb{E}_a(r_a^k|q^k)$ to obtain r_a^k.

3.4 Cost Regret Constraint (***)

Since the budget is limited, we aim to make the best use of the chances for both accuracy improvement and cost savings. In this part, we introduce a penalty term *cost regret*, denoted as $\mathcal{R}_a$, to measure the proportion of costs incurred by incorrect predictions given by the arm a within a budget constraint.

$$\begin{aligned} \mathcal{R}_a &= \frac{\sum_{q\in Q_a^\beta} p(a_k)\|q_a^\beta\|}{\sum_{q\in\{Q_a^\alpha \cup Q_a^\beta\}} p(a_k)\|q_a\|} \\ s.t. & \sum_{q\in\{Q_a^{k-1}\}} p(a_k)\|q_a\| + p(a_k)\|Q_a^k\| \leq \mathcal{B}; \forall a \in \mathcal{A}. \end{aligned} \tag{14}$$

where q_a^β and Q_a^β indicate the failure, i.e., the question and historically assigned questions for arm a that are wrongly answered. $p(a)$ is the unit cost for invoking the LLMs. For black-box LLMs, we calculate $p(a)\|q_a$ as the token-level fees; while for open-sourced LLMs, it will be replaced by the times that we call LLMs. The numerator of $\mathcal{R}_a$ represents the total cost incurred by incorrect answers given by arm a, while the denominator calculates the total cost of all questions answered by arm a. The constraint could effectively ensure that the total cost of selecting arm a for answering questions in the previous $k-1$ iterations and the current iteration k does not exceed a predefined budget $\mathcal{B}$. To control the impacts of over-constraint from *cost regret* which may penalize the model for the costs on necessary trials, we introduce λ as a hyperparameter to control the trade-off between maximizing rewards and minimizing cost regret for a reasonable *exploration* and *exploitation*.

4 Experiments

We conduct experiments on three domain-specific datasets: (i) Commonsense knowledge domain: CommonsenseQA [35]; (ii) Scientific Openbook domain: OpenbookQA [28]; (iii) Medical Domain: MedQA-USMLE [23]. To compare the performance of Coke, we include the baselines from three aspects, i.e., fine-tuned Language Models, KGMs and both API series and local series of LLMs. Additionally, our framework is efficiently runnable on one CPU. To accelerate the matrix computation, we adopt Torch to boost the selection on an NVIDIA GeForce RTX 4090 GPU.

4.1 Datasets

CommonsenseQA [35] (abbreviated as *CSQA*) is a prominent dataset in the commonsense knowledge domain that demands extensive real-world commonsense knowledge. It encompasses 12,102 questions, and ConceptNet [33], one of the largest commonsense knowledge graphs (KG), is frequently employed by existing KGMs for reasoning. Due to the official test set being reserved for leaderboard evaluations, we assess model performance using the in-house (IH) data split as implemented in [25]. **OpenBookQA** [28] (referred to as *OBQA*) is a scientific domain dataset that comprises 5,957 multiple-choice questions from open book exams, each with four options. Answering *OBQA* questions necessitates a comprehensive understanding of fundamental science facts and their applications, which involves understanding scientific principles and applying them to novel situations.

Table 1: Performance comparison among state-of-the-art baselines and Coke on three benchmark datasets in terms of both inferential accuracy and cost saving ($ API fees).

Model	**CommonsenseQA**		**OpenBookQA**		**MedQA**	
	IHdev-Acc.	IHtest-Acc.	Dev-Acc.	Test-Acc.	Dev-Acc.	Test-Acc.
Fine-dtuned Language Models (LMs)						
Bert-base [11]	0.573	0.535	0.588	0.566	0.359	0.344
Bert-large [11]	0.611	0.554	0.626	0.602	0.373	0.367
RoBerta-large [27]	0.731	0.687	0.668	0.648	0.369	0.361
Knowledge Graph Based Small Models (KGMs)						
MHGRN [16]	0.745	0.713	0.786	0.806	-	-
QA-GNN [37]	0.765	0.733	0.836	0.828	0.394	0.381
HamQA [14]	0.769	0.739	0.858	0.846	0.396	0.385
JointLK [34]	0.777	0.744	0.864	0.856	0.411	0.403
GreaseLM [39]	**0.785**	0.742	0.857	0.848	0.400	0.385
GrapeQA [36]	0.782	0.749	0.849	0.824	0.401	0.395
Large Language Models (LLMs) - Local Series						
ChatGLM	0.473	0.469	0.352	0.360	0.346	0.366
Baichuan-7B	0.491	0.476	0.411	0.395	0.334	0.319
Llama2 (7b)	0.561	0.547	0.526	0.466	0.302	0.299
Llama3 (8b)	0.754	0.720	0.762	0.756	0.622	0.691
Coke (Ours)	-	-	-	-	**0.627**	**0.692**
Acc. Imp.% vs. Llama3 (8b)	-	-	-	-	**+ 0.81%**	**+ 0.16%**
Cost Sav. (calls) % vs. Llama3 (8b)	-	-	-	-	**- 1.17%**	**- 0.66%**
Large Language Models (LLMs) - API Series						
GPT3	0.539	0.520	0.420	0.482	0.312	0.289
GPT3.5	0.735	0.710	0.598	0.600	0.484	0.487
GPT-4	0.782	0.802	0.898	0.902	0.739	0.770
Coke (Ours)	**0.802**	**0.824**	**0.902**	**0.908**	**0.746**	**0.778**
Acc. Imp.% vs. GPT-4	**+ 2.56%**	**+ 2.74%**	**+ 1.12%**	**+ 0.67%**	**+ 0.95%**	**+ 1.03%**
Cost Sav. ($) % vs. GPT-4	**- 15.14%**	**- 20.89%**	**- 5.33%**	**- 11.02%**	**- 2.11%**	**- 4.32%**

MedQA-USMLE [23], i.e., *MedQA*, serves as a domain-specific question-answering benchmark focused on medical knowledge. This dataset is derived from the United States Medical Licensing Examination, which is a comprehensive and challenging assessment used to evaluate the competence of prospective medical professionals. MedQA includes a variety of question types that test clinical knowledge, diagnostic reasoning, and medical science applications. The dataset benchmarks the performance in understanding and applying medical knowledge in both healthcare and medicine.

4.2 Baselines

Fine-tuned Language Models abbreviated as LMs. We evaluate our method against standard fine-tuned language models, specifically using Bert-base, Bert-large [11], and RoBerta-large [27].
KGMs We include off-the-shelf small KGMs that integrate KGs for KBQA, including MHGRN [16], QA-GNN [37], HamQA [14], JointLK [34], GreaseLM [39], and GrapeQA [36].
LLMs Our comparison includes two categories of LLMs: the API series (GPT-3, GPT-3.5, and GPT-4) and local series (ChatGLM, Baichuan-7B, Llama2 (7b), and Llama3 (8b).

4.3 Main Results

The overall performance is compared and shown in Table 1. To compare with API series of LLMs, we assemble HamQA, GPT3.5 and GPT-4 as the arms and our proposed Coke has outperformed all four categories of baselines in terms of inferential accuracy. Specifically, we could achieve 2.03%, 0.67% and 1.03% improvements over GPT-4 on three datasets. For the local series of LLMs, we adopt HamQA, Llama2 (7b) and Llama3 (8b). Since Llama2 (7b) and Llama3 (8b) underperform the traditional KGMs with lagging performance on both CSQA and OBQA, we merely conduct the experiments on MedQA with the arm models, i.e., HamQA, Llama2 (7B) and Llama3 (8B). We conclude our observations hereunder. The performance gap among different arms plays a vital role in balancing the accuracy and costs. For example, on CSQA and OBQA, the accuracy of state-of-the-art KGMs are very close to GPT-4 and much better that both GPT 3.5 and local series of LLMs. This facilitates a big improvements on cost saving by invoking more KGMs, where we achieve higher accuracy with much lower costs, i.e., 20.89% and 11.02%. However, on MedQA, where the questions are over-complicated and open-ended for KGMs to infer, there exists a huge performance gap between the best KGM and all the LLMs, especially when compared with GPT-4. Our policy has to rely on sufficient calls of LLMs to ensure accuracy, which indeed increases the costs.

4.4 Hyperparameter Analysis

In this subsection, we conduct a detailed analysis of the important hyperparameters, i.e., λ and $\mathcal{B}$.

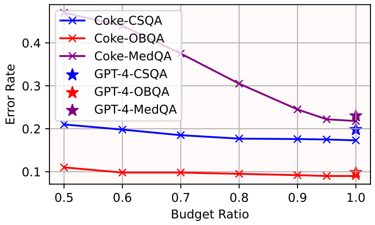


Figure 2: A visualization of Pareto frontier of both inferential accuracy and cost saving as budget $\mathcal{B}$ increases on three datasets.

4.4.1 Budget Study

The Pareto frontier for both inferential accuracy and cost saving is shown in Figure 2. For clear illustration, we replace the accuracy with the error rate which reversely shows the direction of performance changes. To observe when the accuracy and cost reach the balance, we decrease the budget from 1 to 0.5 until `Coke` has a higher error rate than GPT-4 $\mathcal{B}$ ∈{0.5,0.6,0.7...,1}. In this figure, nodes in the lower left positions imply better performance considering both higher accuracy and lower costs. Specifically, our proposed `Coke` could achieve comparable performance to GPT-4 within around 60% budget on both CSQA and OBQA datasets. On MedQA, the slope rapidly drops before the budget ratio reaches 95% and finally outperforms GPT-4 after $\mathcal{B} \geq 0.95$. Since budget $\mathcal{B}$ strictly constrains the opportunities to call LLMs, intuitively, more budgets for LLMs will lead to higher accuracy in most scenarios. However, this does not practically stand for KBQA since the domain knowledge in LLMs is limited. Moreover, purely relying on LLMs will also lose the opportunities to leverage KGMs to further boost the accuracy. On the other hand, more budget will inevitably increase the token-level costs. Consequently, our proposed `Coke` has outperformed GPT-4 on all the datasets with higher accuracy as $\mathcal{B}$ increases. It effectively moves the Pareto frontier for KBQA with LLMs.

4.5 Observations on search of λ

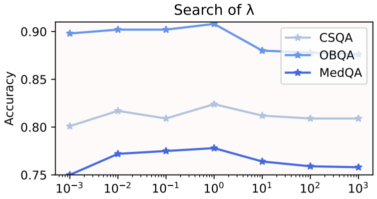


Figure 3: Performance changes based on the search of λ.

The hyperparameter λ is essential for controlling the constraint from cost regret within our framework. We face a dilemma to cautiously decide the value of λ. On one hand, we aim to adopt a large value to strictly penalize models that allocate more resources to failed attempts. On the other hand, an over-penalty will damage the exploration during selection, which means the penalized models will no longer be invoked. Thus, we carefully adjust λ within $\{0.001, 0.1, 1, 10, 100\}$ to modulate the extent to which we minimize cost regret versus maximizing accuracy. The visualization of performance, i.e., inferential accuracy and cost saving, is shown in Figure 3. Consequently, we conclude our observations as follows. A higher λ value signifies a stronger emphasis on cost-efficient decision-making, potentially leading to more conservative model selections. Conversely, a lower λ value may prioritize accuracy over cost savings, resulting in a higher tolerance for resource expenditure on exploration. We adopt the wave peak value of 1 for the final reported performance.

4.6 Ablation Studies

In this part, we investigate the importance of each decomposed expectation in our overall objective of optimization, i.e., $\mathbb{E}_c$ and $\mathbb{E}_a$, as well as the constraint from *cost regret* $\mathcal{R}_a$. The performance of

Table 2: Verification of the importance of $\mathbb{E}_c$, $\mathbb{E}_a$ and $\mathcal{R}_a$ on three datasets.

Ablations	**CommonsenseQA**		**OpenBookQA**		**MedQA**	
	Accuracy	Cost Saving ($)	Accuracy	Cost Saving ($)	Accuracy	Cost Saving ($)
w/o $\mathbb{E}_c$	0.750	**- 47.59%**	0.855	**- 32.10%**	0.607	**- 30.51%**
w/o $\mathbb{E}_a$	0.801	**- 15.42%**	0.880	**- 21.16%**	0.760	**- 1.07%**
w/o $\mathbb{R}_a$	0.800	**- 6.26%**	0.898	**- 3.31%**	0.684	**- 15.98%**
`Coke`	**0.824**	**- 20.89%**	**0.908**	**- 11.02%**	**0.778**	**- 4.32%**

inferential accuracy and cost saving by removing one component are shown in Table 2. Removing $\mathbb{E}_c$ leads to significant reductions in cost savings across all datasets. Specifically, cost savings decrease by 47.59% for CSQA, 32.10% for OBQA, and 30.51% for MedQA. While The exclusion of $\mathbb{E}_a$ results in noticeable reductions in accuracy, especially for MedQA, where accuracy falls to 0.760. Finally, removing $\mathcal{R}_a$ brings decreases in both accuracy and cost savings across all datasets. The observations suggest the unique contribution of each component to the model's overall performance.

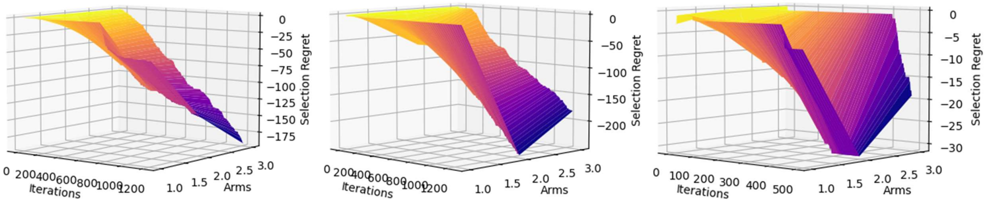


Figure 4: A 3D toy visualization of the selection regret on three datasets as iteration k goes.

4.7 Selection Regret Analysis

Additionally to the proof of the expectation bounds of both $\mathbb{E}_c$ and $\mathbb{E}_a$, we comprehensively evaluate our model selection by visualizing the selection regret in a 3D figure across three datasets in Figure 4. For a clear demonstration of the performance changes, we instantiate the expectation gap $\mathbb{E}[\mu(a_k^*) - \mu(a_k)]$ between best arm a_k^* and the selected arm a_k with a toy example as {-1,0} which indicate the regret of choosing a_k. It sheds light on an intuitive understanding of how regret evolves and converges quickly over iterations, offering valuable insights into the model performance and the correctness of our selection strategy, highlighting the strengths of our proposed method.

4.8 Case Studies

To provide insights into the effectiveness of `Coke` on balancing inferential accuracy and cost saving, we visualize the distribution of model selection on CSQA, OBQA and MedQA in Figure 5, respectively. We could clearly observe the *exploration* process when the color of cubes in the heatmap changes from deep to shallow, e.g., $\{250, 500\}$ intervals on CSQA and MedQA for GPT-4. This makes trials and spends necessary costs on the under-explored model. While the color changes from shallow to deep, e.g., $\{250, 500\}$ intervals on CSQA for HamQA, $\{500, 750\}$ intervals on CSQA and MedQA, $\{100, 200\}$ on OBQA for GPT-4 and $\{750, 1000\}$ intervals for ChatGPT, indicating the *exploitation* process that leverages the best model. The case study sheds light on our superior ability to balance the selection for more accurate and cost-saving candidate models.

5 Related Work

Contemporary research has been focused on generating answers through deductive processes applied to knowledge graphs, as outlined in surveys and studies on the subject [24, 22]. Existing methods predominantly leverage techniques rooted in semantic parsing and information retrieval [31, 26]. For instance, KagNet [25] introduces a schema graph that adeptly captures relational paths among pivotal entities. Nonetheless, these methodologies often delineate the processes of question interpretation and knowledge graph (KG) inference as distinct stages, thereby exposing the models to potential pitfalls associated with the nuanced or implicit facets of query phrasing, including negations and

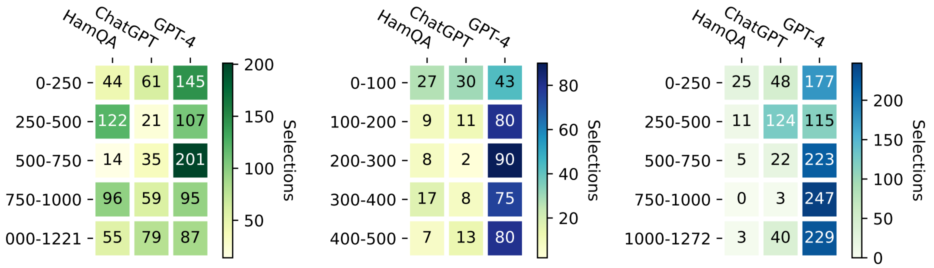


Figure 5: A case study of the model selection on three domain-specific datasets as k goes. The color changes from deep to shallow indicates an *exploration* process, while an *exploitation* reversely.

contextual constraints. Recognizing this limitation, recent innovations have pivoted towards a more integrated approach to reasoning. MHGRN [16] exemplifies this trend by dynamically refining the embeddings of the question context in tandem with the reasoning process, utilizing graph neural networks. This fusion of path encoders with GNNs not only enhances the model's interpretability but also its ability to scale. Building on this, the QA-GNN framework [37] goes further by crafting a work graph wherein the context itself is instantiated as an entity and interlinked with relevant entities through cooccurrence, thereby streamlining the reasoning process. GreaseLM [39] establishes a tighter connection between language models and KGs with a joint training mechanism. While HamQA [14] focuses on combining question understanding and knowledge reasoning, it excels in answering hierarchical questions containing hyponymys.

6 Limitation

Our study faces a major limitation of the distinct performance gaps among existing models considering both inferential accuracy and cost savings, e.g., KGMs and LLMs. This disparity from the underperformance of certain models hinders the overall efficiency of model selection. For example, on MedQA dataset, the gap between the best KGM and GPT-4 is a remarkable 36.50%. This forces our policy to rely on GPT-4 to maintain accuracy, bringing higher costs. However, as a fast-pluggable policy, `Coke` can easily address this concern when more robust models appear to boost the selection.

7 Conclusion

In this paper, we present `Coke`, a novel cost-efficient strategy for LLMs in KBQA while balancing inferential accuracy and cost saving. Our work first formally defines the problem of trading off accuracy and cost for KBQA with LLMs and provides a practical solution for utilizing LLMs in resource-constrained and domain knowledge-required scenarios.`Coke` could effectively integrate two sets of off-the-shelf models, i.e., LLMs and KGMs, and efficiently assign the most promising model for each question within a limited budget by employing a tailored cluster-based Thompson Sampling and a contextual multi-armed bandit. The former models the preliminary selection between LLMs and KGMs based on historical performance, while the latter identifies the best model within a cluster according to question semantics. The overall decision-making is bounded by the cost regret, constraining the selection based on cumulative expenses incurred from model failures. Extensive experiments on three domain-specific benchmark datasets demonstrate the superiority of `Coke` in terms of both inferential accuracy and cost-effectiveness. Our proposed framework could also offer a significantly promising direction for efficient integration of LLMs in various knowledge-based tasks.

Acknowledgement

The work described in this paper was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. PolyU 25208322).

References

[1] Ilaria Amaro, Attilio Della Greca, Rita Francese, Genoveffa Tortora, and Cesare Tucci. Ai unreliable answers: A case study on chatgpt. In *ICHCI*, pages 23–40. Springer, 2023.

[2] Md Adnan Arefeen, Biplob Debnath, and Srimat Chakradhar. Leancontext: Cost-efficient domain-specific question answering using llms. *Natural Language Processing Journal*, page 100065, 2024.

[3] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023.

[4] Hao Chen, Yuanchen Bei, Qijie Shen, Yue Xu, Sheng Zhou, Wenbing Huang, Feiran Huang, Senzhang Wang, and Xiao Huang. Macro graph neural networks for online billion-scale recommender systems. In *Proceedings of the ACM on Web Conference 2024*, pages 3598–3608, 2024.

[5] Hao Chen, Yue Xu, Feiran Huang, Zengde Deng, Wenbing Huang, Senzhang Wang, Peng He, and Zhoujun Li. Label-aware graph convolutional networks. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, page 1977–1980, 2020.

[6] Lingjiao Chen, Matei Zaharia, and James Zou. Frugalgpt: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*, 2023.

[7] Shengyuan Chen, Yunfeng Cai, Huang Fang, Xiao Huang, and Mingming Sun. Differentiable neuro-symbolic reasoning on large-scale knowledge graphs. *Advances in Neural Information Processing Systems*, 36, 2024.

[8] Shengyuan Chen, Qinggang Zhang, Junnan Dong, Wen Hua, Jiannong Cao, and Xiao Huang. Neuro-symbolic entity alignment via variational inference. *arXiv preprint arXiv:2410.04153*, 2024.

[9] Shengyuan Chen, Qinggang Zhang, Junnan Dong, Wen Hua, Qing Li, and Xiao Huang. Entity alignment with noisy annotations from large language models. *arXiv preprint arXiv:2405.16806*, 2024.

[10] Yifang Chen, Chung-Wei Lee, Haipeng Luo, and Chen-Yu Wei. A new algorithm for non-stationary contextual bandits: Efficient, optimal and parameter-free. In *COLT*. PMLR, 2019.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[12] Junnan Dong, Zijin Hong, Yuanchen Bei, Feiran Huang, Xinrun Wang, and Xiao Huang. Clr-bench: Evaluating large language models in college-level reasoning, 2024.

[13] Junnan Dong, Wentao Li, Yaowei Wang, Qing Li, George Baciu, Jiannong Cao, Xiao Huang, Richard Chen Li, and Peter HF Ng. Gradual study advising with course knowledge graphs. In *International Conference on Web-Based Learning*, pages 125–138. Springer, 2023.

[14] Junnan Dong, Qinggang Zhang, Xiao Huang, Keyu Duan, Qiaoyu Tan, and Zhimeng Jiang. Hierarchy-aware multi-hop question answering over knowledge graphs. In *Proceedings of the ACM Web Conference 2023*, pages 2519–2527, 2023.

[15] Junnan Dong, Qinggang Zhang, Huachi Zhou, Daochen Zha, Pai Zheng, and Xiao Huang. Modality-aware integration with large language models for knowledge-based visual question answering, 2024.

[16] Yanlin Feng, Xinyue Chen, Bill Yuchen Lin, Peifeng Wang, Jun Yan, and Xiang Ren. Scalable multi-hop relational reasoning for knowledge-aware question answering. In *EMNLP*, pages 1295–1309, 2020.

[17] Jocelyn Gravel, Madeleine D'Amours-Gravel, and Esli Osmanlliu. Learning to fake it: limited responses and fabricated references provided by chatgpt for medical questions. *Mayo Clinic Proceedings: Digital Health*, 1(3):226–234, 2023.

[18] Zijin Hong, Zheng Yuan, Hao Chen, Qinggang Zhang, Feiran Huang, and Xiao Huang. Knowledge-to-sql: Enhancing sql generation with data expert llm, 2024.

[19] Zijin Hong, Zheng Yuan, Qinggang Zhang, Hao Chen, Junnan Dong, Feiran Huang, and Xiao Huang. Next-generation database interfaces: A survey of llm-based text-to-sql. *arXiv preprint arXiv:2406.08426*, 2024.

[20] Feiran Huang, Zefan Wang, Xiao Huang, Yufeng Qian, Zhetao Li, and Hao Chen. Aligning distillation for cold-start item recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 1147–1157, 2023.

[21] Feiran Huang, Zhenghang Yang, Junyi Jiang, Yuanchen Bei, Yijie Zhang, and Hao Chen. Large language model interaction simulator for cold-start item recommendation. *arXiv preprint arXiv:2402.09176*, 2024.

[22] Xiao Huang, Jingyuan Zhang, Dingcheng Li, and Ping Li. Knowledge graph embedding based question answering. In *WSDM*, pages 105–113, 2019.

[23] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14), 2021.

[24] Yunshi Lan, Gaole He, Jinhao Jiang, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. A survey on complex knowledge base question answering: Methods, challenges and solutions. In *IJCAI*, 2021.

[25] Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. Kagnet: Knowledge-aware graph networks for commonsense reasoning. In *EMNLP-IJCNLP*, pages 2829–2839, 2019.

[26] Lihui Liu, Boxin Du, Jiejun Xu, Yinglong Xia, and Hanghang Tong. Joint knowledge graph completion and question answering. In *KDD*, pages 1098–1108, 2022.

[27] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[28] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*, pages 2381–2391, 2018.

[29] OpenAI. Gpt-4 technical report, 2023.

[30] Diogo Teles Sant'Anna, Rodrigo Oliveira Caus, Lucas dos Santos Ramos, Victor Hochgreb, and Julio Cesar dos Reis. Generating knowledge graphs from unstructured texts: Experiences in the e-commerce field for question answering. In *ASLD@ ISWC*, pages 56–71, 2020.

[31] Apoorv Saxena, Aditay Tripathi, and Partha Talukdar. Improving multi-hop question answering over knowledge graphs using knowledge base embeddings. In *ACL*, pages 4498–4507, 2020.

[32] Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. In chatgpt we trust? measuring and characterizing the reliability of chatgpt. *arXiv preprint arXiv:2304.08979*, 2023.

[33] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. 2016.

[34] Yueqing Sun, Qi Shi, Le Qi, and Yu Zhang. Jointlk: Joint reasoning with language models and knowledge graphs for commonsense question answering. *arXiv preprint arXiv:2112.02732*, 2021.

[35] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. 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 4149–4158, 2019.

[36] Dhaval Taunk, Lakshya Khanna, Siri Venkata Pavan Kumar Kandru, Vasudeva Varma, Charu Sharma, and Makarand Tapaswi. Grapeqa: graph augmentation and pruning to enhance question-answering. In *Companion Proceedings of the ACM Web Conference 2023*, pages 1138–1144, 2023.

[37] Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. Qa-gnn: Reasoning with language models and knowledge graphs for question answering. *NAACL*, 2021.

[38] Qinggang Zhang, Junnan Dong, Hao Chen, Daochen Zha, Zailiang Yu, and Xiao Huang. Knowgpt: Knowledge injection for large language models, 2024.

[39] Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D Manning, and Jure Leskovec. Greaselm: Graph reasoning enhanced language models for question answering. *arXiv preprint arXiv:2201.08860*, 2022.

A Proof of Selection Regret in Cluster-level Thompson Sampling

The expectation bounds of the specific iteration t is:

$$\begin{aligned} SR_t(K) &= \mathbb{E}[R(k)] && (15)\\ &= \mathbb{E}[\mu(a_k^*) - \mu(a_k)] && (16)\\ &= \mathbb{E}_{H_{k-1}}\left[\mathbb{E}[\mu(a_k^*) - \mu(a_k) \mid H_{k-1}]\right] \qquad \text{Note that } a_k \sim a_k^* \text{ when } H_{k-1} \text{ is fixed} && (17)\\ &= \underbrace{\mathbb{E}[U(a_k, H_{k-1}) - \mu(a_k)]}_{\text{Part 1}} + \underbrace{\mathbb{E}[\mu(a_k^*) - U(a_k^*, H_{k-1})]}_{\text{Part 2}} && (18) \end{aligned}$$

According to the properties shown in Eq. 7, the part 2 can be deducted:

$$\begin{aligned} \mathbb{E}[\mu(a_k^*) - U(a_k^*, H_{k-1})] &\leq \mathbb{E}[(\mu(a_k^*) - U(a_k^*, H_{k-1}))^+]\\ &\leq \mathbb{E}\left[\sum_{a_k}^{\mathcal{A}} (\mu(a_k) - U(a_k, H_{k-1}))^+\right]\\ &= \mathbb{E}\left[\sum_{a_k}^{\mathcal{A}} (\mu(a_k) - U(a_k, H_{k-1}))^+\right]\\ &= \sum_{a_k}^{\mathcal{A}} \mathbb{E}\left[(U(a_k, H_{k-1}) - \mu(a_k))^+\right]\\ &\leq k \cdot \frac{\gamma}{\mathcal{K} \cdot k}\\ &= \frac{\gamma}{\mathcal{K}} \end{aligned} \tag{19}$$

Similarly, Eq. 7 suggests that the part 1 is bounded as follows:

$$\begin{aligned} \mathbb{E}[U(a_k, H_{k-1}) - \mu(a_k)] &= \mathbb{E}[2r(a_k, H_{k-1}) + L(a_k, H_{k-1}) - \mu(a_k)]\\ &= \mathbb{E}[2r_c(a_k, H_{k-1})] + \mathbb{E}[L(a_k, H_{k-1}) - \mu(a_k)]\\ \mathbb{E}[L(a_k, H_{k-1}) - \mu(a_k)] &\leq \mathbb{E}[(L(a_k, H_{k-1}) - \mu(a_k))^+]\\ &\leq \mathbb{E}\left[\sum_{a_k}^{\mathcal{A}} (L(a_k, H_{k-1}) - \mu(a_k))^+\right]\\ &= \mathbb{E}\left[\sum_{a_k}^{\mathcal{A}} ((\mu(a_k) - U(a_k, H_{k-1}))^+\right]\\ &= \sum_{a_k}^{\mathcal{A}} \mathbb{E}\left[\mu(a_k) - L(a_k, H_{k-1}))^-\right]\\ &\leq k \cdot \frac{\gamma}{\mathcal{K} \cdot k}\\ &= \frac{\gamma}{\mathcal{K}} \end{aligned} \tag{20}$$

Putting part 1 and Part 2 together, $SR_t = \frac{\gamma}{\mathcal{K}} + \frac{\gamma}{\mathcal{K}} + \mathbb{E}[2r(a_k, H_{k-1})]$. Summing up over t, it can be proved that $SR(\mathcal{K}) \leq 2\gamma + 2\sum_{k=1}^{\mathcal{K}} \mathbb{E}[r(a_k, H_{k-1})]$.

B Broader Impacts

In terms of the potential positive impacts, our work may have impacts on the improved efficiency and better decision-making processes based on LLMs with lower costs in many domain-specific tasks. The potential benefitted scenarios and group of people can be education, medical and healthcare for both research and application. We may also be able to reduce commercial costs for enterprises and innovations in the industry.

This paper does not present any foreseeable negative impacts to society since we focus on the technique of optimizing a model selection policy that is purely serving the KBQA task.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The major contributions to the community have been clearly highlighted in both the Abstract and Introduction. In this paper, we focus on knowledge-based question answering with LLMs. Our moving the Pareto frontier and contributions will definitely inspire the community followed by more meaningful work.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In Section 6, we clearly demonstrate one potential limit of our work brought by the performance gap among existing models. However, as a fast-plugged framework, the limitation can be easily avoided when more robust models appear. We can thereby boost the performance with our superior model selection policy.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: In **Appendix** Section A, we detailedly provide the derivative process of both the upper and lower confidence bounds for expectation $\mathbb{E}_c$.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: To contribute and inspire more valuable research in the community, we have open-sourced our main codes for reproducibility. The codes could be found from this anonymous link: https://anonymous.4open.science/r/NeurIPS-24-Coke-Anonymous-13626/main.py

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: To contribute and inspire more valuable research in the community, we have open-sourced our main codes for reproducibility. The codes could be found from this anonymous link: https://anonymous.4open.science/r/NeurIPS-24-Coke-Anonymous-13626/main.py

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We have detailedly introduced our adopted datasets in Section 4. We also visualize the hyperparameter analysis in Section 4.4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: In our paper, abundant of ablations have been done on all three benchmark domain-specific datasets to showcase the superior performance of our proposed methods. We also provide an optimal seed to control the randomness during the sampling based on Beta distribution.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have demonstrated our used resources for the main results in Section **??**. Indeed, our framework is highly usable with CPU only for general cases as well as commercial scenarios.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We confirm that we have fully complied with the Ethics Policy in this study. We conduct experiments with widely adopted publicly available datasets.The codes are open-sourced for other researchers' further study in the active KBQA community.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have demonstrated out potential positive impacts in **Appendix** Section B. Also this paper does not present any foreseeable negative impacts since we focus on the technique of optimizing a model selection policy that is purely serving the KBQA task.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We merely adopt pre-trained / close-source LLMs on the benchmark knowledge-based QA datasets. The returned answers will only be choices based on careful in-context prompt design. There will be no harmful of risky content generated.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We conduct experiments on widely adopted benchmarks datasets that are publicly available. They are properly cited as the references in our manuscript.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: We strictly follow the main stream of the KBQA community and leverage the existing datasets, as well as their splits with no new datasets used or curated.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: We do not involve any human resources to manually annotate or evaluate.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: We do not involve any human-related experiments in our research. This paper only focuses on the machine comprehension and model selection for knowledge-based question answering.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.