Collaborative Enhancement of Large and Small Models for Question Answering via Dual Knowledge Transfer

Shaofei Wang, Yunan Liu, Xiaolan Tang, Wenlong Chen*

College of Information Engineering, Capital Normal University
Beijing 100048, China
{wangshaofei, chenwenlong}@cnu.edu.cn

Abstract

Our statistical analysis reveals a complementary phenomenon between large language model-based question answering (QA) and small model-based QA. To facilitate dual knowledge transfer between these two paradigms, this paper introduces a collaborative enhancement method of large and small models for question answering. The proposed method consists of two iterative steps: i) **small4large step**, in which the small model first predicts an answer for a given question along with its confidence, and these results are then leveraged as prompts to strengthen the large model's performance; ii) **large4small step**, where the large model enhances the small model through distillation, judgment and reflection. Through iteration of these two steps, the large and small models could enhance each other progressively. Experimental evaluations across eight datasets spanning five domains demonstrate that the proposed method effectively improves the question answering performance of both large and small models simultaneously.

1 Introduction

Question answering (QA) aims to comprehend given questions and predict accurate answers (Kamalloo et al. 2023; Sahoo et al. 2024). QA is a pivotal task in artificial intelligence with broad applications across diverse domains.

Recent advancements in QA primarily fall into two paradigms: i) Small model-based QA. These methods either train task-specific models from scratch using labeled data (Gu et al. 2021; Lee et al. 2020) or fine-tune pre-trained language models (e.g., RoBERTa and GPT-2) (Özkurt 2024; Luo, Luo, and Yang 2024). The strengths of such methods lie in their training efficiency and relatively compact model sizes. However, they heavily rely on large volumes of annotated question-answer pairs. ii) Large model-based QA. This paradigm leverages volume comprehensive knowledge contained in large language models (LLMs) to directly generate answers, while eliminating the need for task-specific training data (Hadi et al. 2023; Kamalloo et al. 2023). While large models faces critical challenges, such as hallucination issues and computational inefficiency during knowledge updating.

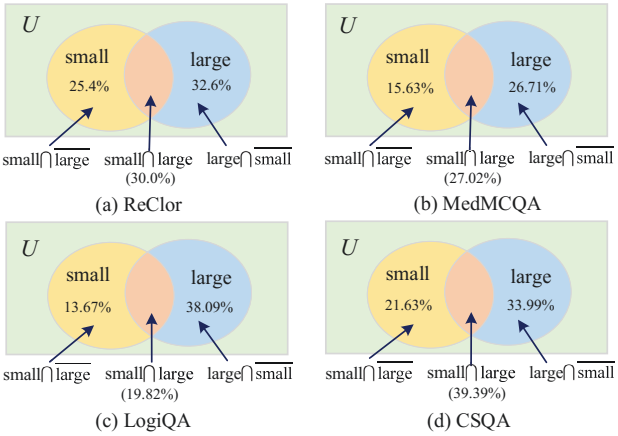


Figure 1: Question answering performance statistics of small and large models across four datasets. The yellow and blue areas represent the proportion of questions answered correctly only by the small model and large model, respectively. These statistics reveal the complementary phenomenon of two models.

To evaluate the performance of large and small models in question answering, we carried out an empirical analysis employing both models on four different QA benchmark datasets[1]. Our study uncovered a **complementary phenomenon** between large and small models in QA tasks.

Complementary Phenomenon: large and small models demonstrate unique strengths in answering questions independently. Some questions can only be answered correctly by the large model, similarly, some can only be answered correctly by the small model. Figure 1 displays the statistical results across four datasets. In the figure, the yellow and blue areas represent the questions answered correctly only by the small model and large model, respectively. The overlapping areas (marked in pink) of two circles represent the questions correctly answered by both models. Take Sub-figure (a) as an example, on the ReClor dataset, 25.4% of questions are answered correctly solely by the small model, 32.6% solely

*Wenlong Chen is the corresponding author.

[1]In the experiment, the small model is the fine-tuned RoBERTa (Liu et al. 2019) using the training data. The large model is LLaMA 3-70B (Dubey et al. 2024).

by the large model, and 30.0% by both models. Motivated by this complementary phenomenon, this study aims to develop an approach that integrates the distinct advantages of both large and small models.

Actually, there have been several approaches aiming to integrate large and small models. However, these methods typically operate in a single-direction manner. For instance, some methods seek to leverage large models to enhance small models (Lee et al. 2023; Guo et al. 2023). Others utilize small models to augment large models (Xu et al. 2024; Zhao et al. 2025). Yet, these methods fail to fully achieve dual knowledge transfer between large and small models.

Therefore, this paper proposes a collaborative enhancement method of large and small models for QA via dual knowledge transfer (DKT). The DKT method contains two iterative steps: i) Small4large step. the small model first predicts an answer along with confidence for a given question. Then a confidence-aware prompting strategy is designed to incorporate the small model's predictions for the large model's reference. ii) Large4small step. the large model enhances the small model via a reflective knowledge distillation mechanism, which consists of three substeps: distillation, judgment, and reflection.

By continuously iterating the small4large and large4small steps, the method progressively improves both large and small models, leveraging each model's strengths to compensate for the other's weaknesses. Experimental results demonstrate that both the large model and the small model exhibit consistent performance improvements across various datasets.

The contributions can be summarized as follows:

- We propose a collaborative dual knowledge transfer method for large and small models in QA, enabling bidirectional enhancement by leveraging one model's strengths to compensate for the other's weaknesses.
- The method comprises two iterative stages: i) Small4large step. the small model enhances the large model through a confidence-aware prompting strategy. ii) Large4small step. the large model improves the small model via a reflective knowledge distillation mechanism. Experimental results validate the effectiveness of the proposed method.

2 Related Work

The related work in this paper can be categorized into the following two aspects:

Small and large models for QA. QA tasks mainly contain text-based QA, web-based QA, and information retrieval-based QA (Mervin 2013; Caballero 2021). In this paper, we focus on the text-based QA, in which given a textual question, and the model is asked to predict the answer based on its internal knowledge (Nassiri and Akhloufi 2023; Chaturvedi, Pandit, and Garain 2018).

Early work mainly focused on either training QA models from scratch or fine-tuning small models with labeled QA data (Liu et al. 2019; Nassiri and Akhloufi 2023). Among these approaches, fine-tuning small models using domain-specific data has emerged as a crucial research direction, such as BioBERT (Lee et al. 2020) and Lawformer (Xiao et al. 2021). These models provide high training and knowledge updating efficiency, yet they depend heavily on large amounts of labeled QA pairs.

Subsequently, LLMs have developed stronger reasoning and comprehension abilities, and can be applied to QA tasks (Yue 2025; Kamalloo et al. 2023). To enhance the performance of LLM-based QA, some approaches employ supervised fine-tuning (SFT) to adjust the parameters of large models using domain-specific QA data (Bolton et al. 2024; Labrak et al. 2024). However, these SFT-based models typically demand significant computational resources for training. Alternatively, some methods leverage retrieval-augmented generation (RAG) to improve large models without modifying their original parameters (He et al. 2024; Sharma et al. 2024; Sen, Mavadia, and Saffari 2023; Jiang et al. 2024). This approach combines the strengths of parametric knowledge in LLMs with non-parametric retrieval from external knowledge sources. As previously discussed in this paper, there exists a complementary phenomenon between small and large models. In contrast to the aforementioned approaches, this work introduces the DKT method to integrate small and large models.

Collaboration of small and large models. There exist several approaches that integrate small and large models to achieve superior question answering performance (Chen et al. 2024; Xu et al. 2024; Zhao, Zhang, and Zong 2023). These approaches mainly contain two categories: i) leveraging large models to enhance small models (Hsieh et al. 2023; Ye et al. 2022). For example, LLM-labeling (Wang et al. 2021) leverages the large models to label training data for small models to reduce the cost of training. RLAIF (Lee et al. 2023) utilizes the feedback of the large model as reward to improve small models. CombinedQA (Guo et al. 2023) incorporates large models to rewrite existing questions and simultaneously generates new questions to enrich the medical training data; ii) utilizing small models to augment large models (Chen et al. 2024; Wang et al. 2025). SuperICL (Xu et al. 2024) injects the results of small models into the large model's prompt as in-context information. RoSe (Zhao et al. 2025) introduces reference answers for large models through a role guidance strategy to alleviate the over-rely on reference answers. These methods typically adopt a single-direction manner, where either large models are employed to strengthen small models or vice versa. Different from these methods, our proposed DKT method implements dual enhancement by leveraging both small4large and large4small knowledge transfer mechanisms.

3 Method Description

Figure 2 presents the overview of the proposed method, which encompasses two key steps: i) the **small4large** step, in which the small model improves the performance of the large model through a confidence-aware prompting strategy (Section 3.1); ii) the **large4small** step, the large model enhances the capabilities of the small model via a reflective knowledge distillation with three substeps: distillation, judgment, and reflection (Section 3.2). By continuously iterating

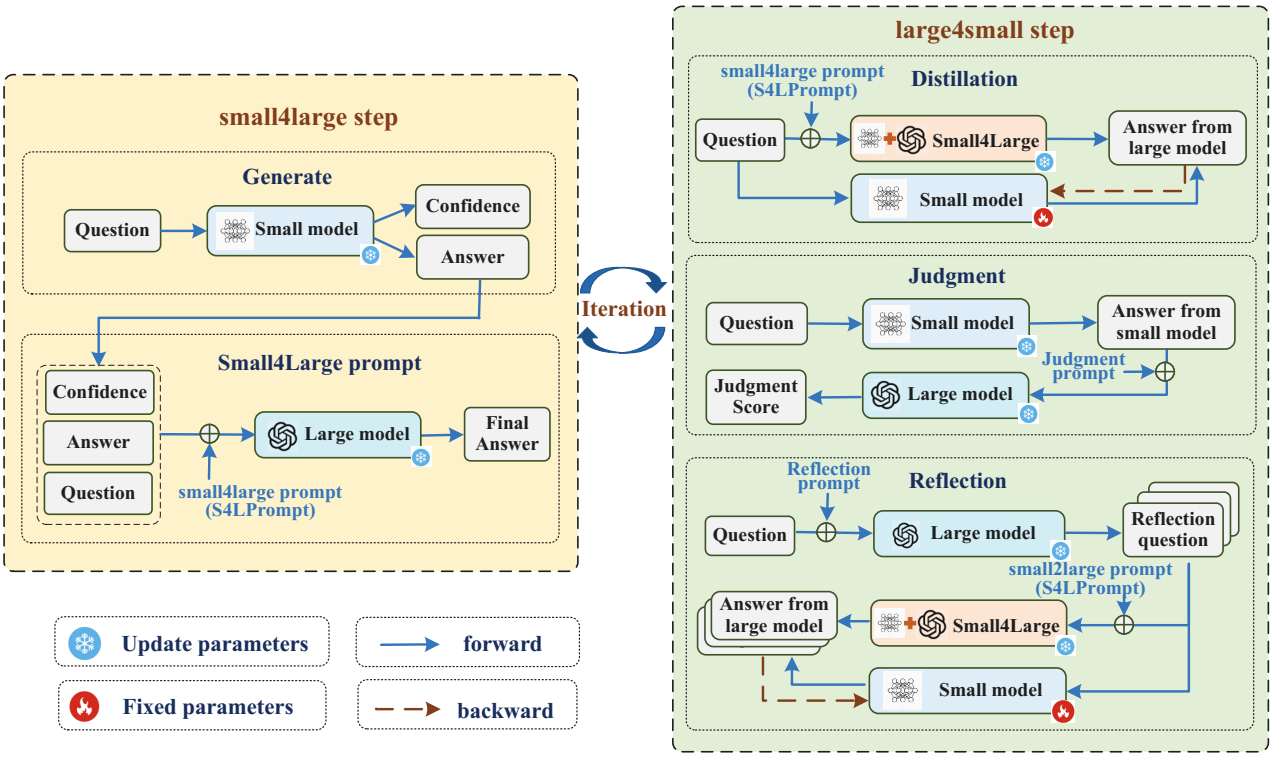


Figure 2: The framework of DKT method, which contains two steps: i) **small4large** step, in which the small model improves the performance of the large model through a confidence-aware prompting strategy (left); ii) **large4small** step, the large model enhances the capabilities of the small model via a reflective knowledge distillation mechanism (right).

through the small4large and large4small steps, both the large and small models can progressively achieve better results.

3.1 Small4large Step

The goal of the small4large step is to leverage small models to enhance large models. Given the huge number of parameters in large models, their parameters remain unchanged in proposed method. Instead, small models are employed to improve the responses of large models through prompting. Meanwhile, given that small models always underperform large models, thus we take into account the confidence of the small model when constructing the prompt. Only when the confidence of a prediction exceeds a predefined threshold, we use the prediction as prompt for large models.

The procedure of small4large step is shown in left parts of Figure 2. Specifically, for a given question q, the small model first predicts an answer a_{small} along with its probability p_{small} (confidence).

$$a_{\text{small}}, p_{\text{small}} = \text{argmax}\, p(a|q;\theta) = \text{argmax} f_{\text{small}}(q;\theta) \quad (1)$$

where a_{small} represents the output of the small model and p_{small} denotes its output probability. Then the generated result and confidence of the small model are used as prompts and fed into the large model as follows:

$$a_{\text{large}} = \begin{cases} f_{\text{large}}\left(\text{S4LPrompt}(q, a_{\text{small}}, p_{\text{small}})\right), & \text{if } p_{\text{small}} > p_0, \\ f_{\text{large}}\left(\text{S4LPrompt}(q)\right), & \text{otherwise.} \end{cases} \quad (2)$$

as shown in the equation 2, if p_{small} exceeds a predetermined threshold p_0, we utilize the output from the small model to enhance the performance of the large model using S4LPrompt [2].

3.2 Large4small Step

Unlike the small4large step, the large4small step focuses on enhancing small models by leveraging knowledge from small4large step. Given a small model, the parameters of small models are fine-tuned through a reflective distillation process. Algorithm 1 illustrates the specific procedure of the large4small step. As depicted in the algorithm, the large4small step comprises three main substeps: distillation, judgment, and reflection. Note that for the given training dataset D, we partition it into distillation data D_{distill} and judgment data D_{judge}, i.e. $D = D_{\text{distill}} \cup D_{\text{judge}}$, which are used in the distillation and judgment substeps respectively.

Substep 1: Distillation. In the distillation substep, the small model (student) learns from the small4large step (teacher). Given a question-answer pair $(q_d, a_d) \in D_{\text{distill}}$, where each pair consists of a question q_d and its ground-truth

[2]The S4LPrompt is *"You are a question answering assistant. Please carefully read the question and predict the most likely correct answer. Question: {q}. My classmate believes the answer is {a_{small}}. Your task is to respond to the question above. Of course, you can take your classmate's answer as a reference if you think it's correct, or you can disregard it if you believe otherwise."*

Algorithm 1: Large4Small step

Input: large model, initialized small model θ, QA dataset $D = D_{\text{distill}} \cup D_{\text{judge}}$

Output: Optimized small model θ_{new}

while not converged **do**
  substep 1: Distillation
    Update small model θ by distilling small4large step (teacher) on dataset $D_{distill}$ with equation 3
  substep 2: Judgment
    Get predictions a_{small} from the small model for each query q_j in D_{judge}
    Compute judgment scores $s_{\text{large}}(a_{\text{small}})$ through the large model (equation 4)
    Construct rejected dataset D_{reject}
  substep 3: Reflection
    Construct the reflection dataset $D_{\text{reflection}}$ based on the rejected dataset D_{reject} with equation 5
    Update small model θ by distilling small4large step (teacher) on dataset $D_{\text{reflection}}$ with equation 6
end while
Return small model θ_{new}

answer a_d, we first input the query q_d into the small4large step (teacher) and derive the output a_{large}, then the small model not only learn from the gold target answer a_d but also the output a_{large} from small4large step as follows:

$$\begin{aligned}\mathcal{L}_{\text{KD}} =&(1-\lambda)\cdot \sum_{(q_{\text{d}},a_{\text{d}})\in D_{\text{distill}}} logP(a_{\text{d}} \mid q_{\text{d}};\theta)+ \\ &\lambda \cdot \sum_{q_{\text{d}}\in D_{\text{distill}}} logP(a_{\text{large}} \mid q_{\text{d}};\theta)\end{aligned} \tag{3}$$

where θ represents the parameters of the small model.

Substep 2: Judgment. In the judgment substep, we leverage the large model to judge the predictions output by the small model and construct a rejected dataset D_{reject}, which comprises samples where the small model demonstrates suboptimal performance.

Specifically, we first input each question $q_j \in D_{\text{judge}}$ into the small model and generate the small model's prediction a_{small}. The large model then evaluates the quality of this prediction by computing a score as follows:

$$s_{\text{large}}(a_{\text{small}}) = f_{\text{large}}\left(\text{JudgePrompt}(q_j, a_{\text{small}})\right) \tag{4}$$

$s_{\text{large}}(a_{\text{small}})$ measures how well the small model's output a_{small} aligns with the large model's judgment [3].

[3]The JudgmentPrompt can leverage the large model to judge the small model. The specific prompt is *" As a judge assistant, your primary responsibility is to to judge the correctness of the small model's predicted answer to the given question on a scale from 0 to 1, where: 0 indicates a completely incorrect or irrelevant response, 1 indicates a perfectly accurate and complete answer. Here is the input: question: $\{q\}$, predicted answer: $\{a_{small}\}$. Output: Provide a numerical score (from 0 to 1) reflecting the answer's correctness."*

If this score falls below a predefined threshold s_0, i.e., $s_{\text{large}}(a_{\text{small}}) < s_0$, the sample (q_j, a_{small}) is deemed inadequately handled by the small model and added to the rejected dataset D_{reject}. By traversing over all samples in D_{judge}, we could construct the rejected dataset $D_{\text{reject}} \subseteq D_{\text{judge}}$.

Substep 3: Reflection. Rather than directly distilling knowledge from D_{reject} to the small model, the reflection substep first expands D_{reject} to reflection samples D_{reflect} , which are then utilized to optimize the small model. Our motivation stems from the observation that reflection samples could generate analogous instances to those in D_{reject} and make the small model generalize beyond limited examples.

Specifically, for each rejected sample $q_r \in D_{\text{reject}}$, we generate k different reflection samples via:

$$D_{\text{reflect}} = \bigcup_{q_{\text{r}}\in D_{\text{reject}}} f_{\text{large}}\left(\text{ReflectPrompt}(q_r)\right), \tag{5}$$

This process expands D_{reject} by leveraging the large model's capacity to generate analogous instances through the reflection prompt and ultimately constructs the augmented reflection dataset D_{reflect}[4].

We then leverage the reflection dataset D_{reflect} to distill knowledge from the large model into the small model. Since reflection samples lack ground-truth answers, we update the small model solely with the teacher model's outputs as follows:

$$\mathcal{L}_{\text{reflect-KD}} = \sum_{q_{\text{r}}\in D_{\text{reflect}}} \log P(a_{\text{large}} \mid q_r;\theta) \tag{6}$$

where a_{large} is the output from large model given q_r.

Discussion. The proposed DKT method can be viewed as a collaborative bootstrapping approach. For the small model, it first enhances the large model through a prompting strategy (small4large), which, in turn, reciprocally strengthens the small model itself through the large4small process:

$$\text{Loop 1: small} \xrightarrow[\text{small4large}]{\text{prompting}} \text{large} \xrightarrow[\text{large4small}]{\text{distillation}} \text{small} \tag{7}$$

For the large model, it improves the small model through reflective distillation, while the small model further self-enhances via prompt-based refinement:

$$\text{Loop 2: large} \xrightarrow[\text{large4small}]{\text{distillation}} \text{small} \xrightarrow[\text{small4large}]{\text{prompting}} \text{large} \tag{8}$$

[4]The ReflectionPrompt is *"You serve as an AI Assistant specialized in data generation tasks. Please review the provided question $\{q_r\}$ and generate K new questions that follow a similar structure. The newly created questions should encompass two categories of analogous examples: i) Semantic Consistency with Varied Descriptions: These questions should retain the same underlying meaning as the original but be phrased differently. ii) Semantically Distinct yet Conceptually Related: These questions should explore a different topic but maintain a similar question format or logical structure. For example, if the provided question is 'What is the capital of China?', an analogous question in the first category could be 'Which city serves as China's capital?', while an example from the second category could be 'What is the capital of France?' ".*

Dataset	Train	Valid	Test
MedQA	20,356	2,544	2,546
MedMCQA	117,949	2,816	2,816
JEC-QA	12,642	4,215	4,215
SciQ	11,679	1,000	1,000
CSQA	8,520	1,221	1,221
ProtoQA	44,975	52	52
LogiQA	7,376	651	651
ReClor	4,138	500	500

Table 1: Statistics of datasets in the experiments.

4 Experiments

4.1 Experimental Settings

Datasets. The experiments are conducted on eight datsets across five domains (see Table **??**): MedQA (Jin et al. 2021) and MedMCQA (Pal, Umapathi, and Sankarasubbu 2022) in medical domain, JEC-QA (Zhong et al. 2020) in legal domain, SciQ (Welbl, Liu, and Gardner 2017) in science QA, CSQA (Talmor et al. 2019) and ProtoQA (Boratko et al. 2020) in commonsense domain, and LogiQA (Liu et al. 2021) and ReClor (Yu et al. 2020) in logical domain. Among them, ProtoQA is a generative question answering task, while the others are multiple-choice question answering tasks. Since the test sets of certain datasets remain unavailable, this study follows the approach adopted in prior work (Kumar et al. 2024; Liu et al. 2023) by presenting the results on the validation set and samples an equal number of instances from the training set to serve as the validation set.

Evaluation Metrics. For multiple-choice QA task, accuracy is utilized as the evaluation metric, and for generative QA task, we utilize Max Answers@10 as the evaluation metric. Max Answers@10 could measure the coverage of model predictions over ground-truth answers when restricted to returning at most 10 answers (Boratko et al. 2020).

Comparison Methods. We compared recent state-of-the-art methods as follows:

- For the scenario of enhancing small models with large models, we conducted a comparative study of three methods: LLM-labeling (Wang et al. 2021), RLAIF (Lee et al. 2023), and CombinedQA (Guo et al. 2023).
- For the scenario of leveraging small models to enhance large models, we compare with the following two methods: RoSe (Zhao et al. 2025) and SuperICL (Xu et al. 2024).

During the comparative evaluation, we employ a diverse set of small and large models.

For **small models**, we employ BioBERT (Lee et al. 2020) (medical domain) and Lawformer (Xiao et al. 2021) (legal domain). Additionally, we utilize Qwen2.5-1.5B (Qwen 2024) for ProtoQA tasks and a fine-tuned version of RoBERTa (Liu et al. 2019) for other datasets.

For **large models**, we integrate three large models with varying parameter sizes: LLaMA 3-70B (Dubey et al. 2024), DeepSeek r1-dislled-Qwen-14B (abbreviated as DeepSeek r1-14B) (Guo et al. 2025), and QwQ-32B (Qwen 2025).

Experimental Setup. In the small4large step, the confidence threshold p_0 is set as 0.6 for multiple-choice tasks, and 0.1 for generative QA task. For multiple-choice tasks, the confidence score is defined as the predicted probability of the selected option. In contrast, for generative QA tasks, we compute the confidence score by averaging the token-level probabilities of the generated answer sequence. In the judgment substep, the quality score threshold s_0 is 0.8. The number of iterations for small4large and large4small is 2. The reflection sample number $k = 3$. When predicting, the temperature of large models is set to 0.6, and when generating the reflection samples, the temperature is set to 1.0. All hyperparameters are determined based on the validation set.

4.2 Main Results

Table 2 presents the answer quality of different methods when applied to eight QA datasets with different large and small models. Specifically, Lines 5-11 illustrate the results when the large model is LLaMA 3-70B, Lines 12-18 display the performance with the large model being DeepSeek r1-14B, and Lines 19-25 showcase the results for QwQ-32B. Based on these results, it is evident that for eight QA datasets, the proposed method DKT can enhance the performance of both small and large models simultaneously.

Results of Large4small. We first analyze how large models enhance the performance of small models. As can be seen from the table, with the augmentation from large models, all methods are able to improve the performance of small models, and the proposed method achieves the most significant enhancements. For instance, on the CSQA dataset for commonsense reasoning, the small model initially attains an accuracy of 60.69%. After applying the proposed DKT method, its accuracy is boosted to 63.96%, 64.29%, and 64.54% respectively. Moreover, the proposed method also demonstrates superior performance compared to other approaches that leverage large models to enhance small models, such as LLM-labeling, RLAIF, and CombinedQA.

From the results, we can observe that on the vast majority of datasets, the performance of large models surpasses that of small models. In such scenarios, leveraging large models to enhance small models can yield favorable performance. Furthermore, we find that even in cases where the large model's performance is inferior to that of the small model, the proposed method remains effective. For example, for commonsense question-answering task ProtoQA, the Max Answers@10 of small model is 49.1, while that of DeepSeek r1-14B is 42.6. Through dual knowledge transfer of the proposed method, the performance of small model is improved to 51.7, thereby validating the effectiveness of our approach.

Results of Small4large. In the small4large step where a small model is employed to enhance a large model, the proposed method DKT has demonstrated its capacity to improve the large model's performance. In most cases, the results of small models are inferior to those of large models. However, in the DKT method, the proposed confidence-aware prompting strategy first filter the predictions of small models by their corresponding confidence. Then, these results with higher confidence are introduced into the prompt for the large model to reference when answering questions.

#	Model	Medical		Legal	Science	Common		Logical	
		MedQA	MedMCQA	JEC-QA	SciQ	CSQA	ProtoQA	LogiQA	ReClor
1	Small model	33.07	42.65	24.22	63.0	60.69	49.1	33.49	55.4
2	LLaMA 3-70B	52.40	53.73	38.98	83.5	73.38	38.2	57.91	62.6
3	DeepSeek r1-14B	68.34	62.18	40.85	94.5	81.90	42.6	62.37	71.4
4	QwQ-32B	81.74	71.20	58.10	94.4	86.57	40.2	78.80	94.6
	Large Model= LLaMA 3-70B								
	large4small								
5	LLM-labeling	33.31	42.68	24.39	63.2	62.08	49.4	33.18	55.6
6	RLAIF	33.11	42.79	24.93	64.0	62.57	46.7	33.79	55.4
7	CombinedQA	33.62	43.00	25.05	64.1	62.98	49.1	34.72	55.6
8	DKT-small (Ours)	**34.84**	**44.11**	**25.91**	**65.8**	**63.96**	**50.7**	**36.25**	**58.0**
	small4large								
9	RoSe	52.00	54.01	38.93	83.4	72.81	**40.2**	57.91	63.6
10	SuperICL	52.67	54.65	39.07	83.1	72.07	39.1	58.22	64.2
11	DKT-large (Ours)	**54.71**	**56.92**	**40.28**	**83.8**	**74.61**	**40.2**	**61.14**	**66.0**
	Large Model= DeepSeek r1-14B								
	large4small								
12	LLM-labeling	33.62	42.97	24.63	63.9	62.74	48.6	33.95	56.4
13	RLAIF	33.74	43.25	24.65	64.5	62.90	47.3	34.41	56.0
14	CombinedQA	33.94	44.03	25.10	65.4	63.14	49.4	35.33	57.0
15	DKT-small (Ours)	**35.90**	**44.92**	**27.33**	**66.2**	**64.29**	**51.7**	**37.63**	**59.2**
	small4large								
16	RoSe	68.54	62.39	41.57	94.4	82.23	43.1	62.98	71.8
17	SuperICL	68.81	62.61	41.30	**94.7**	82.56	**44.0**	64.06	72.2
18	DKT-large (Ours)	**69.40**	**63.25**	**42.09**	**94.7**	**82.80**	43.8	**64.52**	**73.4**
	Large Model= QwQ-32B								
	large4small								
19	LLM-labeling	34.05	43.29	24.72	64.2	63.14	48.3	33.95	56.6
20	RLAIF	34.13	43.36	24.67	64.8	63.23	48.4	34.10	56.2
21	CombinedQA	34.49	44.39	24.86	65.5	63.55	49.1	36.25	58.2
22	DKT-small (Ours)	**36.02**	**45.13**	**27.31**	**66.0**	**64.54**	**51.3**	**37.79**	**60.0**
	small4large								
23	RoSe	81.81	71.66	59.34	94.0	86.90	40.8	78.80	94.0
24	SuperICL	82.05	**72.30**	58.20	**94.5**	87.22	41.3	78.96	94.6
25	DKT-large (Ours)	**82.13**	72.27	**59.57**	94.4	**87.39**	**41.6**	**79.57**	**94.8**

Table 2: The experimental results on various small and large models. For the ProtoQA task, we report the Max Answers@10 metric, while accuracy is reported for other tasks. The best results are highlighted in bold.

Finally, the performance of large models can be improved. For example, the performance of the large model LLaMA 3-70B on ReClor is 62.6%, and the proposed method can boost this performance to 66.0%.

4.3 Ablation Study

In this section, we conducted an ablation study to evaluate the impact of four key components in DKT: i) iteration between small4large and large4small, ii) confidence-aware prompting in the small4large, iii) distillation substep employed in the large4small, iv) judge and reflection substeps utilized in the small4large. Experimental results on the valid sets of four datasets are presented in Table 3. The large model used in these experiments is DeepSeek r1-14B, and the small model specifications are detailed in Section 4.1.

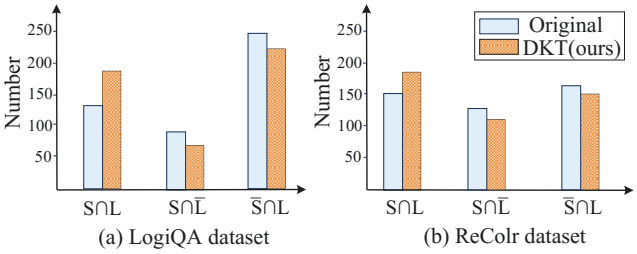
From the experimental results, we observe that removing either components step leads to a decline for both large and small models. For instance, for the large model, when the confidence-aware prompting strategy is absent, the response accuracy of the large model drops from 73.6% to 71.8% on ReClor. When the distillation in large4small step is removed, the accuracy of small model decreases from 59.6% to 57.6%. As for the substeps judgment and reflection, when these substeps are removed, the accuracy of small model drops from 59.6% to 57.8%. These results demonstrate the importance of each component.

4.4 Statistics of Complementary Phenomenon

As mentioned previously, large and small models exhibit a complementary phenomenon. That is, some questions can only be correctly answered by the large models, and some can only be answered accurately by the small models. Our proposed DKT method is capable of enhancing dual knowledge transfer between large and small models. To this end,

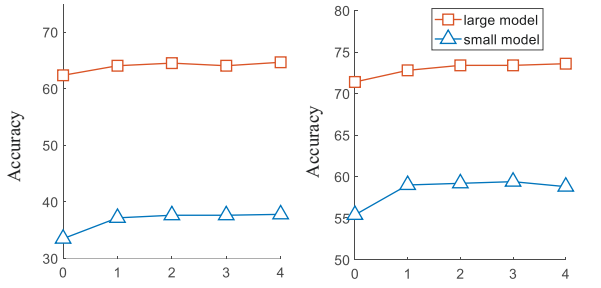
models	Small Model				Large Model			
	LogiQA	ReClor	CSQA	MedMCQA	LogiQA	ReClor	CSQA	MedMCQA
DKT (ours)	37.33	59.6	64.46	45.06	64.67	73.6	82.8	63.35
w/o Iteration	37.17	59.0	64.05	44.92	64.06	72.8	82.56	63.32
w/o Confidence prompt	36.41	58.2	63.80	44.42	62.98	71.8	82.15	62.54
w/o Distillation substep	34.25	57.6	62.82	43.57	63.44	72.8	82.56	63.10
w/o Judge and reflection substeps	36.41	57.8	63.64	44.67	63.59	72.4	82.47	63.21
Small model	33.33	56.0	60.93	42.72	-	-	-	-
Large model	-	-	-	-	62.83	71.2	81.98	62.43

Table 3: Ablation study to evaluate the impact of four key components in the DKT method: i) iteration between small4large and large4small, ii) confidence-aware in the small4large, iii) distillation subset in the large4small, iv) judge and reflection subsets in the small4large.



(a) LogiQA dataset (b) ReColr dataset

Figure 3: Statistical results illustrating the complementary phenomenon observed in the LogiQA dataset (a) and the Recolr dataset (b). $S \cap L$ denotes questions where both models answer correctly, $S \cap \overline{L}$ denotes questions where only the small model answers correctly, and $\overline{S} \cap L$ denotes questions where only the large model answers correctly.



(a) LogiQA dataset (b) ReClor dataset

Figure 4: Iteration results for large4small and small4large knowledge transfer. The left presents outcomes on the LogiQA dataset, and the right displays results from the ReClor dataset.

we have conducted a statistical analysis of the following three subsets of data: i) questions where both models answer correctly (denoted as $S \cap L$), ii) questions where only the large model answers correctly (denoted as $\overline{S} \cap L$), and iii) questions where only the small model answers correctly (denoted as $S \cap \overline{L}$).

Figure 3 reports the statistical results obtained from the LogiQA dataset (a) and the Recolr dataset (b). In these experiments, the small model employed is a fine-tuned RoBERTa, and the large model is DeepSeek r1-14B. From the results, we can see that the proposed DKT method can reduce the numbers of questions in $\overline{S} \cap L$ and $S \cap \overline{L}$, while increasing the number of questions in $S \cap L$. On the ReClor dataset, the number of questions that both large small models answered correctly has increased from 150 to 178 through the method DKT. Meanwhile, the number of questions correctly answered by the small model increased from 277 to 290, and that of the large model from 313 to 330. The statistics demonstrate that the DKT method has achieved dual knowledge transfer and enhance the accuracy provided by both models simultaneously.

4.5 Iteration Between Large4small and Small4large

In the proposed DKT method, we need to iterate through two sub-steps: large4small and small4large. In this section, we analyze the impact of the number of iterations of these two sub-steps (large4small and small4large) on performance. Figure 4 presents the experimental results, from which it can be observed that as the iterations progress, the performance of both large and small models gradually improves in the first two iterations. After two epochs of iteration, the best results are achieved, and as the iterations continue, the performance gradually converges.

5 Conclusion

To alleviate the complementary phenomenon between small and large models, as well as the limitations of existing single directional enhancement methods, we propose a DKT method to achieve dual enhancement between small and large models. The DKT method consists of two components: i) small4large step: The small model's predictions, accompanied by their confidences, are incorporated as prompts to enhance the large model. ii) large4small step: The small model is improved through a reflective knowledge distillation mechanism, which comprises three substeps: distillation, judgment, and reflection. Experiments on diverse QA datasets demonstrate the effectiveness of DKT, particularly in leveraging the complementary strengths of small and large models to achieve accuracy improvements for both.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant 6257077228, by the Beijing Digital Education Research Project under Grant BDEC2024ZX065, and by the Beijing Outstanding Youth Talent Development Program under Grant BPHR202203118.

References

Bolton, E.; Venigalla, A.; Yasunaga, M.; Hall, D.; Xiong, B.; Lee, T.; Daneshjou, R.; Frankle, J.; Liang, P.; Carbin, M.; et al. 2024. Biomedlm: A 2.7 b parameter language model trained on biomedical text. *arXiv preprint arXiv:2403.18421*.

Boratko, M.; Li, X.; O'Gorman, T.; Das, R.; Le, D.; and Mccallum, A. 2020. ProtoQA: A Question Answering Dataset for Prototypical Common-Sense Reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 1122–1136.

Caballero, M. 2021. A brief survey of question answering systems. *International Journal of Artificial Intelligence & Applications*, 12(5).

Chaturvedi, A.; Pandit, O.; and Garain, U. 2018. CNN for text-based multiple choice question answering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 272–277.

Chen, D.; Zhuang, Y.; Zhang, S.; Liu, J.; Dong, S.; and Tang, S. 2024. Data shunt: Collaboration of small and large models for lower costs and better performance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 11249–11257.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; and Poon, H. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1): 1–23.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Guo, Z.; Wang, P.; Wang, Y.; and Yu, S. 2023. Improving small language models on pubmedqa via generative data augmentation. *arXiv preprint arXiv:2305.07804*.

Hadi, M. U.; Qureshi, R.; Shah, A.; Irfan, M.; Zafar, A.; Shaikh, M. B.; Akhtar, N.; Wu, J.; Mirjalili, S.; et al. 2023. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints*.

He, X.; Tian, Y.; Sun, Y.; Chawla, N.; Laurent, T.; LeCun, Y.; Bresson, X.; and Hooi, B. 2024. G-retriever: Retrieval-augmented generation for textual graph understanding and question answering. *Advances in Neural Information Processing Systems*, 37: 132876–132907.

Hsieh, C.-Y.; Li, C.-L.; YEH, C.-K.; Nakhost, H.; Fujii, Y.; Ratner, A. J.; Krishna, R.; Lee, C.-Y.; and Pfister, T. 2023. Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*.

Jiang, B.; Wang, Y.; Luo, Y.; He, D.; Cheng, P.; and Gao, L. 2024. Reasoning on Efficient Knowledge Paths: Knowledge Graph Guides Large Language Model for Domain Question Answering. In *2024 IEEE International Conference on Knowledge Graph*, 142–149. IEEE Computer Society.

Jin, D.; Pan, E.; Oufattole, N.; Weng, W.-H.; Fang, H.; and Szolovits, P. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14): 6421.

Kamalloo, E.; Dziri, N.; Clarke, C.; and Rafiei, D. 2023. Evaluating Open-Domain Question Answering in the Era of Large Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 5591–5606.

Kumar, P. S.; et al. 2024. Bridging the Knowledge Gap: Improving BERT models for answering MCQs by using Ontology-generated synthetic MCQA Dataset. In *The International FLAIRS Conference Proceedings*, volume 37.

Labrak, Y.; Bazoge, A.; Morin, E.; Gourraud, P.-a.; Rouvier, M.; and Dufour, R. 2024. BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Lee, H.; Phatale, S.; Mansoor, H.; Mesnard, T.; Ferret, J.; Lu, K.; Bishop, C.; Hall, E.; Carbune, V.; Rastogi, A.; et al. 2023. RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. *arXiv preprint arXiv:2309.00267*.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234–1240.

Liu, H.; Ning, R.; Teng, Z.; Liu, J.; Zhou, Q.; and Zhang, Y. 2023. Evaluating the logical reasoning ability of chatgpt and gpt-4. *arXiv preprint arXiv:2304.03439*.

Liu, J.; Cui, L.; Liu, H.; Huang, D.; Wang, Y.; and Zhang, Y. 2021. LogiQA: a challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 3622–3628.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Luo, X.; Luo, Y.; and Yang, B. 2024. Question answering systems based on pre-trained language models: Recent progress. In *International Conference on Intelligent Information Processing*, 173–189. Springer.

Mervin, R. 2013. An overview of question answering system. *International Journal Of Research In Advance Technology In Engineering*, 1: 11–14.

Nassiri, K.; and Akhloufi, M. 2023. Transformer models used for text-based question answering systems. *Applied Intelligence*, 53(9): 10602–10635.

Özkurt, C. 2024. Comparative Analysis of State-of-the-Art Q&A Models: BERT, RoBERTa, DistilBERT, and ALBERT on SQuAD v2 Dataset. *Chaos and Fractals*, 1(1): 19–30.

Pal, A.; Umapathi, L. K.; and Sankarasubbu, M. 2022. MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. In *Proceedings of the Conference on Health, Inference, and Learning*, Proceedings of Machine Learning Research, 248–260.

Qwen, T. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Qwen, T. 2025. QwQ-32B: Embracing the Power of Reinforcement Learning.

Sahoo, P.; Singh, A. K.; Saha, S.; Jain, V.; Mondal, S.; and Chadha, A. 2024. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications. *arXiv preprint arXiv:2402.07927*.

Sen, P.; Mavadia, S.; and Saffari, A. 2023. Knowledge Graph-augmented Language Models for Complex Question Answering. In *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations*, 1–8.

Sharma, S.; Yoon, D. S.; Dernoncourt, F.; Sultania, D.; Bagga, K.; Zhang, M.; Bui, T.; and Kotte, V. 2024. Retrieval augmented generation for domain-specific question answering. *arXiv preprint arXiv:2404.14760*.

Talmor, A.; Herzig, J.; Lourie, N.; and Berant, J. 2019. 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*, 4149–4158.

Wang, S.; Liu, Y.; Xu, Y.; Zhu, C.; and Zeng, M. 2021. Want To Reduce Labeling Cost? GPT-3 Can Help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 4195–4205.

Wang, Z.; Wang, Z.; Le, L.; Zheng, S.; Mishra, S.; Perot, V.; Zhang, Y.; Mattapalli, A.; Taly, A.; Shang, J.; Lee, C.-Y.; and Pfister, T. 2025. Speculative RAG: Enhancing Retrieval Augmented Generation through Drafting. In *The Thirteenth International Conference on Learning Representations*.

Welbl, J.; Liu, N. F.; and Gardner, M. 2017. Crowdsourcing Multiple Choice Science Questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, 94–106.

Xiao, C.; Hu, X.; Liu, Z.; Tu, C.; and Sun, M. 2021. Lawformer: A pre-trained language model for chinese legal long documents. *AI Open*, 2: 79–84.

Xu, C.; Xu, Y.; Wang, S.; Liu, Y.; Zhu, C.; and McAuley, J. 2024. Small Models are Valuable Plug-ins for Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*, 283–294.

Ye, J.; Gao, J.; Li, Q.; Xu, H.; Feng, J.; Wu, Z.; Yu, T.; and Kong, L. 2022. ZeroGen: Efficient Zero-shot Learning via Dataset Generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 11653–11669.

Yu, W.; Jiang, Z.; Dong, Y.; and Feng, J. 2020. ReClor: A Reading Comprehension Dataset Requiring Logical Reasoning. In *International Conference on Learning Representations*.

Yue, M. 2025. A survey of large language model agents for question answering. *arXiv preprint arXiv:2503.19213*.

Zhao, L.; Wang, Y.; Liu, Q.; Wang, M.; Chen, W.; Sheng, Z.; and Wang, S. 2025. Evaluating large language models through role-guide and self-reflection: A comparative study. In *The Thirteenth International Conference on Learning Representations*.

Zhao, Y.; Zhang, J.; and Zong, C. 2023. Transformer: A general framework from machine translation to others. *Machine Intelligence Research*, 20(4): 514–538.

Zhong, H.; Xiao, C.; Tu, C.; Zhang, T.; Liu, Z.; and Sun, M. 2020. JEC-QA: a legal-domain question answering dataset. In *Proceedings of the AAAI conference on artificial intelligence*, 9701–9708.