Multi-stage Rationale Generation and Contrastive Learning for Chain-of-Thought Distillation

Juan Li, Wen Zhang*, Mingchen Tu, Shijian Li*
Zhejiang University
Hangzhou, China
{lijuan18, zhang.wen, mingchentz, shijianli}@zju.edu.cn

***Abstract*—Recently, knowledge distillation (KD) has gained significant attention for its ability to transfer stronger reasoning capabilities from large language models (LLMs) to small language models (SLMs). Among the KD methods, chain of thought distillation (CoTD) has been explored widely, distilling the intermediate rationales of LLMs to bolster the reasoning performance of SLMs. Existing CoTD works use LLMs as teachers to generate rationales and fine-tune student SLMs based on correct rationales, neglecting the crucial value of incorrect rationales. To address the issue, we propose a novel CoTD framework MGCL that leverages Multi-stage rationale Generation and dual-level Contrastive Learning to enhance the reasoning ability of SLMs by improving the quality of rationales. Specifically, we first employ a self-correction pipeline to achieve multi-stage generation, which diagnoses errors and synthesizes refined and high-quality rationales. Then, we apply a dual-level contrastive learning mechanism to align sub-questions and rationales across different stages. Finally, the SLM is optimized via multi-task joint learning. Extensive experiments demonstrate the superiority of MGCL on commonsense and biomedical reasoning tasks.**

***Index Terms*—Large language model, Knowledge distillation, Chain-of-thought distillation**

I. Introduction

Large Language Models (LLMs) have exhibited impressive performance in complex reasoning tasks, attributed to chain-of-thought (CoT) prompting [1]. However, the prohibitive parameter scale of LLMs poses challenges for practical deployment. While small language models (SLMs) offer superior efficiency, they have limited reasoning ability due to their constrained model capacity. To bridge this gap, chain-of-thought distillation (CoTD) methods have been proposed to transfer reasoning capability from LLMs to SLMs.

Previous CoTD methods, such as Std-CoT [2], follow the paradigm that first uses an LLM to generate rationales, and then fine-tunes the SLM on the correct generated rationales. To enrich the fine-tuning data, Fine-tune-CoT [3] generates multiple rationales for each training sample. SCoTD [4] samples and filters multiple and diverse rationales. TinyLLM [5] generates rationales from multiple teachers. Subsequent CoTD methods primarily focus on designing sophisticated fine-tuning objectives. CasCoD [6] proposes two cascading learning steps, including a rationale and an answer learning step to alleviate the impact of spurious correlations between questions and answers. DSS [7] investigates the mutual relations of the two tasks and formulates it as maximizing the mutual information of the representation features of the two tasks. Additionally, knowledge-augmented distillation has emerged, for example, KARD [8] supplements distillation with external knowledge retrieval and a neural reranker. TinyThinker [9] highlights the risk of imitation in CoTD methods, proposing a coarse-to-fine knowledge internalization process with self-reflection to foster genuine integration of reasoning and underlying knowledge.

Despite the success of existing CoTD methods, the quality of the distilled rationales is often overlooked, which may lead to the propagation of reasoning errors from LLMs to SLMs. Recently, the work [10] investigates the importance of rationale quality and selects high-quality rationales via GPT-4o [11] based on domain-specific knowledge for distillation in mental health applications. However, it relies on high-quality external knowledge bases, which are often difficult to acquire.

To address this issue, we propose a novel CoTD framework MGCL, distilling knowledge into the SLMs via Multi-stage rationale Generation and dual-level Contrastive Learning. Specifically, for multi-stage rationale generation, we first obtain the initial sub-questions via question decomposition and the initial rationales given the question and its sub-questions. The rationales are classified as correct if the answer is correct, otherwise they are labeled as erroneous. Then we generate correct rationales for the questions that have wrong rationales, and extract knowledge by comparing the correct and wrong rationales of the same question. Subsequently, we refine the sub-questions with the knowledge and the rationales with the refined sub-questions. For dual-level contrastive learning, we define contrastive loss considering the initial and refined sub-questions as well as the initial and refined rationales. Finally, we optimize a multi-task joint learning objective that consists of answer prediction task, rationale prediction task, and dual-level contrastive learning. Our MGCL is more flexible as it achieves rationale refinement without any external knowledge source. Experiments on the commonsense and biomedical reasoning tasks verify the effectiveness of our method.

Our contributions are as follows:

- We introduce a CoTD framework that focuses on improving the quality of rationales without external knowledge.
- We propose multi-stage rationale generation to ensure the high quality of the rationales, and dual-level contrastive learning to jointly optimize the SLMs.
- We conduct experiments on commonsense and biomedical reasoning tasks, verifying the effectiveness of MGCL.

*Corresponding author.

II. Related Work

A. Knowledge Distillation

Recent advancements in LLMs are hindered by significant computational demands and deployment complexities, posing challenges for real-world applications. To mitigate resource demands in deployment, knowledge distillation (KD) has emerged as a pivotal paradigm and demonstrated efficacy across diverse applications. KD techniques train smaller student models to mimic the behavior of larger teacher models. Early approaches primarily align output distributions or intermediate representations, for example, [12] employs Kullback-Leibler divergence (KLD) loss to match the predicted and the ground-truth probability distributions. MiniLLM [13] replaces the forward KLD objective with reverse KLD to prevent the student model from overestimating the low-probability regions of the teacher distribution. DISTILLM [14] proposes an adaptive off-policy approach designed to enhance the efficiency in utilizing student-generated outputs.

B. Chain-of-Thought Distillation

Despite the emergent reasoning capabilities elicited by chain-of-thought prompting in LLMs, their deployment remains challenging. This has fueled the rise of CoT distillation, which empowers SLMs by leveraging the logical reasoning processes of LLMs. For example, Std-CoT [2] fine-tunes the SLM by utilizing the correct rationales generated from LLMs. Fine-tune-CoT [3] proposes to enrich the diversity of the fine-tuning data by generating multiple rationales for each sample and TinyLLM [5] mitigates the limited knowledge diversity by learning from multiple teachers. To alleviate spurious correlations between questions and answers, Distilling step-by-step [15] and CasCoD [6] train small models through both answer and rational prediction tasks. Moreover, KARD [8] and D&R Distillation [16] introduce external knowledge bases to improve the performance of knowledge-intensive tasks. SKIntern [17] guides SLMs to gradually absorb symbolic knowledge via a linear decay schedule under curriculum learning. NesyCD [18] distinguishes general and specific knowledge. It first fine-tunes the student model via general distillation, and subsequently analyzes errors to generate knowledge that augments the neural distillation process. While these methods focus on knowledge augmentation, they often ignore the potential of contrastive signals between initial and refined logic, which our MGCL framework exploits.

III. Method

We propose the framework MGCL that consists of multi-stage rationale generation, dual-level contrastive learning, and multi-task joint optimization. Multi-stage rationale generation aims to improve the quality of rationales, and dual-level contrastive learning considers the logical consistency between the initial and refined sub-questions and rationales. Considering answer and rationale prediction and dual-level contrastive learning, the student model is optimized by a multi-task joint learning objective. The framework is shown in Fig. 1.

A. Multi-stage Rationale Generation

The quality of rationales has been proven important for knowledge transfer [10]. To enhance the quality of rationales, we propose a multi-stage rationale generation module to learn from errors and obtain the refined rationales. The module consists of three steps: sub-question and rationale initialization, error-driven knowledge extraction, and knowledge-guided sub-question and rationale refinement.

a) Sub-question and Rationale Initialization: We first design a question decomposition prompt P_d to guide an LLM T to decompose complex questions into ordered and logically progressive sub-questions. The process is formalized as:

$$C = T(Q, P_d) \tag{1}$$

where Q and C respectively indicate the original question and the initial decomposed sub-questions. Then we use an LLM to generate the rationale and answer with the question Q, the initial sub-questions C, and options O as input. The process is defined as:

$$R^I, A^I = T(Q, C, O, P_r) \tag{2}$$

where R^I and A^I respectively indicate the output rationale and answer. P_r is the prompt for generating the rationale and answer. The generated rationale is considered as correct (R^+) when the answer is correct, otherwise it is labeled as incorrect (R^-). Thus, the rationale set can be defined as $R^I = R^+ \cup R^-$. Following this criterion, we respectively partition the questions Q and sub-questions C into corresponding subsets $\{Q^+, Q^-\}$ and $\{C^+, C^-\}$, based on the correctness of their associated rationales and answers.

b) Error-Driven Knowledge Extraction: Considering the incorrect rationales, we propose to extract knowledge from errors and further guide the LLM to generate better rationales. Specifically, we first generate the correct rationale for the questions in Q^-, the generated correct rationale is as follows:

$$R^{-+} = T(Q^-, C^-, A, P_a) \tag{3}$$

where P_a is the prompt given the correct answer A as input. Then we perform error diagnosis and obtain the knowledge by comparing the reasoning of the incorrect rationale and the correct rationale of the same question. For a specific question $q_i \in Q^-$, the knowledge is defined as:

$$K_{q_i} = T(q_i, R^-_{q_i}, R^{-+}_{q_i}, P_k) \tag{4}$$

where P_k is the prompt for knowledge extraction. The extracted knowledge should explain the reasoning error and provide the key basis and correct approach for the revision. $R^-_{q_i} \in R^-$ and $R^{-+}_{q_i} \in R^{-+}$ represent the wrong and correct rationale of q_i, and K_{q_i} is the knowledge of q_i.

c) Knowledge-Guided Sub-question and Rationale Refinement: With the extracted knowledge, we rectify the sub-questions and rationales contrasting with the initial incorrect sub-questions and rationales. We design the prompts P_s and

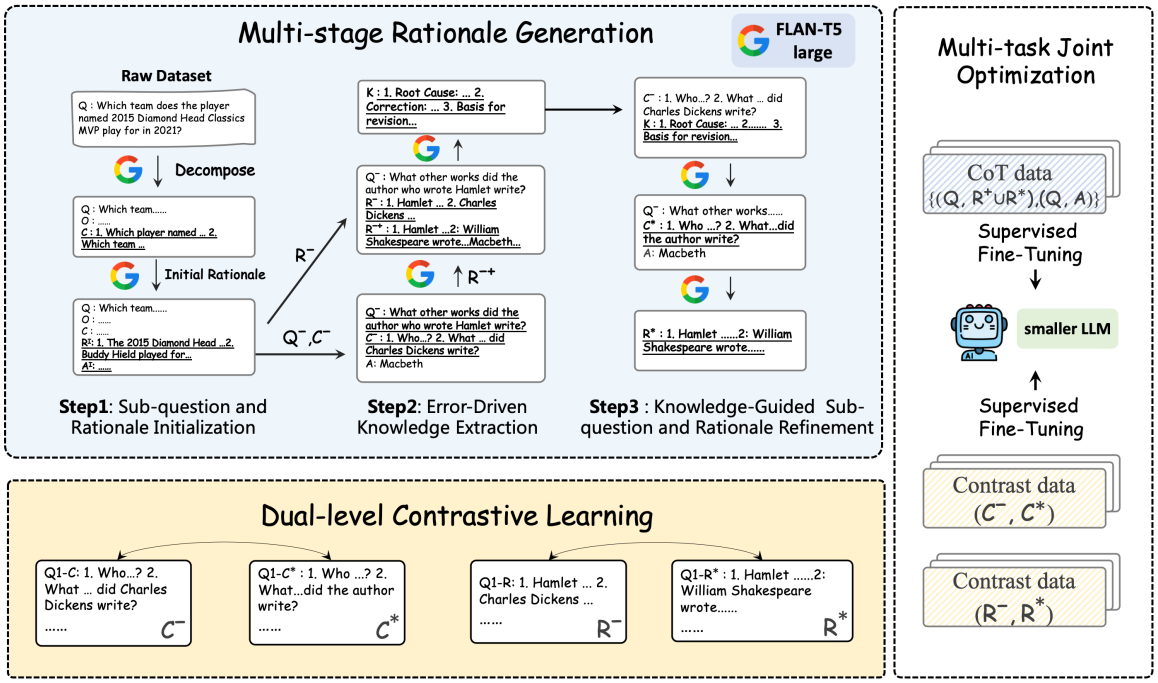

Fig. 1. The overall architecture of the proposed MGCL framework. It consists of three modules: multi-stage rationale generation, dual-level contrastive learning, and multi-task joint optimization. The terms C and C^* indicate the initial and refined sub-questions, R and R^* indicate the initial and refined rationales. Underlined terms in the multi-stage generation module denote the outputs from the preceding process.

P_t to rectify the sub-questions and rationales, respectively. The refinement is as follows:

$$C^* = T(C^-, K, P_s) \tag{5}$$

$$R^* = T(Q^-, C^*, A, P_t) \tag{6}$$

where C^* and R^* are the refined sub-questions and rationales. The final rationales used for training the student model are thus defined as $R_f = R^+ \cup R^*$, which means we use the initial correct rationales for the question in Q^+ and the refined rationales for the question in Q^-.

B. Dual-level Contrastive learning

As learning from the rationale and answer may lead the small models to mimic the behavior of teacher models, we apply dual-level contrastive learning to distinguish the logical inconsistency. Specifically, we design the dual-level contrastive loss $\mathcal{L}_s(C^-, C^*)$ to contrast the initial and refined sub-questions, and $\mathcal{L}_t(R^-, R^*)$ to contrast the initial and refined rationales. The two loss terms are calculated as:

$$\mathcal{L}_s(C, C^*) = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\mathbf{c}_i^\top \mathbf{c}_i^*/\tau)}{\sum_{j=1}^{N}\exp(\mathbf{c}_i^\top \mathbf{c}_j^*/\tau)} \tag{7}$$

$$\mathcal{L}_t(R, R^*) = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\mathbf{r}_i^\top \mathbf{r}_i^*/\tau)}{\sum_{j=1}^{N}\exp(\mathbf{r}_i^\top \mathbf{r}_j^*/\tau)} \tag{8}$$

where N denotes the batch size, $\mathbf{c}_i \in \mathbb{R}^d$ and $\mathbf{c}_i^* \in \mathbb{R}^d$ represent the representations of the initial sub-questions c_i and the refined sub-questions c_i^*, respectively. Similarly, $\mathbf{r}_i$ and $\mathbf{r}_i^*$ denote the representations of the initial rationale r_i and the refined rationale r_i^*. τ is the temperature parameter. For each anchor $\mathbf{c}_i$, the corresponding refined representation $\mathbf{c}_i^*$ serves as the positive sample, while all other refined representations $\{\mathbf{c}_j^*\}_{j\neq i}$ within the same batch act as negative samples. The overall contrastive learning loss is defined as:

$$\mathcal{L}_C = \mathcal{L}_s(C^-, C^*) + \mathcal{L}_t(R^-, R^*) \tag{9}$$

C. Multi-Task Joint Optimization

We employ the rationales and answers as supervision to train the student model. To further enhance the model's ability to discriminate between valid and invalid sub-questions and rationales, we integrate the dual-level contrastive loss into a unified multi-task learning objective. The overall loss function $\mathcal{L}$ integrates three complementary components: answer prediction, rationale generation, and dual-level contrastive learning, which is defined as follows:

$$\mathcal{L} = \alpha\mathcal{L}_A + \beta\mathcal{L}_R + \lambda\mathcal{L}_C \tag{10}$$

where $\mathcal{L}_A$, $\mathcal{L}_R$ are cross-entropy losses for answer prediction and reasoning chain generation, respectively. $\mathcal{L}_C$ is dual-level contrastive loss. α, β and λ are hyper-parameters. For answer prediction loss $\mathcal{L}_A$, we train the student model to predict the correct answer. For reasoning chain generation loss $\mathcal{L}_R$, we train the student to generate high-quality rationales.

IV. Experiment

A. Experimental Setting

a) Datasets: We perform experiments on commonsense reasoning and biomedical reasoning tasks. For commonsense

TABLE I
ACCURACY(%) ON SIX REASONING DATASETS. WE USE THE T5-LARGE MODEL AS THE TEACHER MODEL, AND USE T5-SMALL AND T5-BASE AS STUDENT MODELS. BASELINE RESULTS ARE COPIED FROM TINYLLM [5].

Setting	Method	Commonsense Reasoning				Biomedical Reasoning	
		OBQA	ARC	PIQA	RIDDLE	PubmedQA	BioASQ
Teacher	T5-large	38.60	40.94	61.81	18.04	53.75	62.60
T5-small	Inference	16.60	19.31	20.78	13.33	38.00	47.97
	PINTO	46.40	26.87	48.10	25.29	60.00	80.49
	LoRA	37.80	27.12	39.93	39.80	53.75	78.05
	Full Fine-tuning	41.60	27.47	42.33	42.75	56.25	78.86
	Standard KD	45.80	29.53	49.29	36.27	58.00	81.30
	Distill-step-by-step	46.40	30.47	50.38	36.67	59.00	81.30
	TinyLLM	49.40	33.05	53.65	**51.18**	62.00	85.37
	MGCL	**50.60**	**34.08**	**58.00**	**51.18**	**64.25**	**86.18**
T5-base	Inference	31.00	23.00	30.47	30.78	48.00	57.72
	PINTO	50.40	38.63	52.12	34.90	61.75	82.93
	LoRA	51.40	37.25	47.66	53.14	62.00	82.93
	Full Fine-tuning	56.60	38.88	47.55	52.55	64.75	89.43
	Standard KD	55.40	43.69	55.93	42.94	64.25	86.18
	Distill-step-by-step	56.80	43.86	56.37	45.69	64.75	86.18
	TinyLLM	64.20	48.50	60.17	60.78	66.25	90.24
	MGCL	**64.80**	**48.60**	**66.49**	**61.57**	**68.00**	**91.06**

reasoning, we use OBQA [19], ARC [20], PIQA [21], and RIDDLE [22] as benchmarks. For biomedical reasoning, we adopt PubmedQA [23] and BioASQ [24] datasets.

b) Baselines: Following TinyLLM [5], we compare our MGCL with inference-only, fine-tuning, and knowledge distillation methods. Inference-only means to leverage the pre-trained model for evaluation without training. For fine-tuning methods, we consider LoRA [25], full fine-tuning, and PINTO [26]. For KD methods, the baselines standard KD [12] fine-tunes the student models to mimic the teacher's answers, Distill-step-by-step [15] leverages rationales, and TinyLLM obtains rationales from multiple teachers.

c) Implementations: We apply FLAN-T5 large (T5-large) as the teacher LLM to annotate the rationales, FLAN-T5 small (T5-small) and base (T5-base) are used as the student SLM. For the multi-stage rationale generation module, the temperature is set as 0.7. For dual-level contrastive learning, we set temperature as $\{1.0, 0.1, 0.01\}$. The weight parameters α and β are set as $\{1.0, 0.5, 0.1\}$, and λ is set as $\{0.01, 0.1\}$. The learning rate is set to $\{$5e-4, 5e-5, 5e-6$\}$ and the optimizer is AdamW. We set batch size to 8 or 16, max input length to 512, and steps to $\{1000, 2000, 5000\}$.

B. Main Results

We conduct experiments across six diverse datasets. The detailed results are summarized in Table I.

Our observations are as follows: (1) MGCL consistently outperforms all baseline methods across different student model scales, validating that MGCL can provide higher-quality supervision for the student models. (2) Notably, MGCL achieves the most significant performance gain on the PIQA dataset, where the best performance surpasses the best-performing baseline TinyLLM by margins of 4.35% and 6.32% when using T5-small and T5-base as the student model, respectively. Given that PIQA requires physical commonsense reasoning, these gains underscore MGCL's ability to capture logical dependencies that are often lost in vanilla distillation.(3) Remarkably, MGCL achieves better reasoning accuracy compared to the teacher model T5-large. Taking PubmedQA as an example, MGCL applied to T5-small (64.25%) and T5-base (68.00%) can both exceed the reasoning accuracy of the T5-large teacher (53.75%). The phenomenon validates that our method does not merely mimic the teacher but actively rectifies teacher errors.

TABLE II
ABLATION RESULTS ON DIFFERENT COMPONENTS FOR PIQA AND PUBMEDQA DATASETS.

Setting	T5-small		T5-base	
	PIQA	PubmedQA	PIQA	PubmedQA
w/o RP	57.45	61.50	66.59	65.25
w/o SCL	56.80	60.00	65.94	68.50
w/o RCL	57.02	63.00	65.07	67.25
MGCL	58.00	64.25	66.49	68.00

C. Ablation Study

We perform an ablation study to dissect the individual contributions of each component within the MGCL framework on two representative benchmarks PIQA and PubmedQA. We evaluate three variants by removing the rationale generation loss $\mathcal{L}_R$ (w/o RP), the sub-question contrastive loss $\mathcal{L}_s$ (w/o

TABLE III
IMPACT OF TRAINING DATA SCALE ON REASONING ACCURACY FOR PIQA AND PUBMEDQA DATASETS ACROSS DIFFERENT STUDENT MODEL SIZES.

Setting	T5-small		T5-base	
	PIQA	PubmedQA	PIQA	PubmedQA
20%	53.97	59.50	61.04	63.25
40%	54.95	59.00	64.20	64.25
60%	53.43	61.50	63.76	66.00
80%	54.41	61.25	65.51	66.25
100%	58.00	64.25	66.49	68.00

TABLE IV
PERFORMANCE COMPARISON OF STUDENT MODELS DISTILLED FROM DIFFERENT TEACHER MODELS.

Setting	T5-large		GPT-4o-mini	
PubmedQA	53.75		76.5	
	T5-large		GPT-4o-mini	
	T5-small	T5-base	T5-small	T5-base
PubmedQA	64.25	68.00	62.50	68.50

SCL), or the rationale contrastive loss $\mathcal{L}_t$ (w/o RCL) to show the impact of these components. As illustrated in Table II, MGCL consistently outperforms most variants on PIQA and PubmedQA in the T5-small and T5-base setting, demonstrating that each module can enhance the reasoning capabilities of SLMs across diverse scales. In the student model T5-small, the performance drop observed in the variants without SCL and RCL underscores that contrastive supervision is pivotal for small models to distinguish between valid and invalid reasoning logics. While in the student model T5-base, the variant slightly exceeds the full MGCL without RP on PIQA or without SCL on PubmedQA. This may be attributed to the sufficient internal logic as the model scale increases. Moreover, PIQA involves physical commonsense where the student model may prioritize direct answers over explicit generative explanations, explaining the stability of the "w/o RP" variant on the T5-base model. In contrast, PubmedQA involves knowledge-intensive reasoning and requires rigorous evidence-based deduction.

D. Model analysis

a) Impact of the size of fine-tuning data: Following previous studies, we investigate the impact of fine-tuning data scale on model performance. Specifically, we vary the proportion of the training data for answer and rationale prediction, while keeping the entire contrastive data constant to evaluate how the scaling of primary supervision signals affects the reasoning accuracy. We use PIQA and pubmedqa as examples and vary the ratio as {20%, 40%, 60%, 80%, 100%}. The results are summarized in Table III. From the table, we observe a consistent upward trend in reasoning accuracy as the volume of fine-tuning data increases across both student model scales T5-small and T5-base. For instance, T5-base improves from 63.25% to 68.00% as the data ratio scales from 20% to 100%. It indicates that the higher-quality rationales provided by our framework serve as effective supervisory signals, allowing the models to continuously extract and internalize reasoning logics. T5-base models show a more consistent and substantial growth on PubmedQA compared to PIQA. This suggests that knowledge-intensive domains require higher data coverage to navigate complex evidence. While T5-small exhibits minor fluctuations, reflecting the instability of the smaller models when modeling higher-order logic.

b) Impact of the teacher model in fine-tuning: We use T5-large and GPT-4o-mini as the teacher model to explore the impact of the teacher model in CoT distillation. From Table IV, we observe that although GPT-4o-mini exhibits superior performance compared to T5-large on PubmedQA, it does not yield a correspondingly superior student model. This discrepancy may be attributed to the fact that GPT-4o-mini's reasoning often incorporates a vast amount of external commonsense, which are difficult for a small model with limited parameters to internalize. In contrast, the teacher model T5-large, which shares the same architecture as the student, provides rationales that are more logically digestible. This indicates that a substantial capacity gap between the teacher and student can indeed impede effective knowledge transfer.

E. Case study

As presented in Table V, we conduct case studies to validate that MGCL can generate higher-quality rationales. The two cases are respectively from the PubmedQA and PIQA datasets. Taking the second question as an example, the initial wrong rationale contained irrelevant facts such as "a banana is a fruit" and meaningless assertions like "the batter is only enough for one loaf of bread". In contrast, the refined rationale lists core ingredients, clarifies the operational logic, and supplements critical operational details demonstrating the model's ability to generate practical, task-relevant guidance.

V. CONCLUSION

In this paper, we propose the chain-of-thought distillation framework MGCL to generate higher quality of rationales generated by the large language model for fine-tuning the small language model. MGCL introduces three modules, including multi-stage rationale generation, dual-level contrastive learning, and multi-task joint optimization. For multi-stage rationale generation, we first initialize sub-question via question decomposition, and initialize rationales based on the question and sub-questions. Secondly, we generate rationales given the correct answer and extract knowledge for erroneous rationales. Thirdly, we refine the sub-questions and rationales with the extracted knowledge. Furthermore, we design dual-level contrastive learning to constrain the initial and refined sub-questions and rationales. Subsequently, we jointly optimize the answer and rationale prediction, as well as the dual-level contrastive learning. Experimental results on commonsense reasoning and biomedical reasoning tasks demonstrate the effectiveness of MGCL.

TABLE V
COMPARISON BETWEEN INITIAL WRONG RATIONALES AND MGCL REFINED RATIONALES.

Question	Wrong Rationale	Refined Rationale
We analyzed the pharmacokinetic-pharmacodynamic relationship of vancomycin to determine the drug exposure parameters that correlate with the efficacy and nephrotoxicity of vancomycin in patients with methicillin-resistant Staphylococcus aureus pneumonia and evaluated the need to use peak concentration in therapeutic drug monitoring (TDM)..... Is peak concentration needed in therapeutic drug monitoring of vancomycin?	1. Vancomycin is effective for MRSA. 2. Vancomycin is not well metabolized by the nephrotoxic side of S. aureus. 3. Peak concentrations are not important.	1. C_{min}/MIC and AUC_{0-24}/MIC correlate with efficacy. 2. Safety is predicted by C_{min} and AUC_{0-24}. Step 3. Peak concentration is not among the significant predictors.
How to bake banana bread	1. A banana is a fruit. 2. To make banana bread, you place rolled out dough in a greased loaf pan. 3. The batter is only enough for one loaf of bread.	1. You will need flour, sugar, baking soda and salt. 2. Roll out the dough in a rolling pin. 3. Place dough in the oven for 30 minutes.

ACKNOWLEDGMENT

This work is founded by National Natural Science Foundation of China (No. NSFC62306276), Yongjiang Talent Introduction Programme (2022A-238-G), and Fundamental Research Funds for the Central Universities (226-2023-00138).

REFERENCES

[1] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *NeurIPS*, 2022.

[2] L. C. Magister, J. Mallinson, J. Adámek, E. Malmi, and A. Severyn, "Teaching small language models to reason," in *ACL (2)*. Association for Computational Linguistics, 2023, pp. 1773–1781.

[3] N. Ho, L. Schmid, and S. Yun, "Large language models are reasoning teachers," in *ACL (1)*. Association for Computational Linguistics, 2023, pp. 14 852–14 882.

[4] L. H. Li, J. Hessel, Y. Yu, X. Ren, K. Chang, and Y. Choi, "Symbolic chain-of-thought distillation: Small models can also "think" step-by-step," in *ACL (1)*. Association for Computational Linguistics, 2023, pp. 2665–2679.

[5] Y. Tian, Y. Han, X. Chen, W. Wang, and N. V. Chawla, "Beyond answers: Transferring reasoning capabilities to smaller llms using multi-teacher knowledge distillation," in *WSDM*. ACM, 2025, pp. 251–260.

[6] C. Dai, K. Li, W. Zhou, and S. Hu, "Improve student's reasoning generalizability through cascading decomposed cots distillation," in *EMNLP*. Association for Computational Linguistics, 2024, pp. 15 623–15 643.

[7] X. Chen, H. Huang, Y. Gao, Y. Wang, J. Zhao, and K. Ding, "Learning to maximize mutual information for chain-of-thought distillation," in *ACL (Findings)*, ser. Findings of ACL, vol. ACL 2024. Association for Computational Linguistics, 2024, pp. 6857–6868.

[8] M. Kang, S. Lee, J. Baek, K. Kawaguchi, and S. J. Hwang, "Knowledge-augmented reasoning distillation for small language models in knowledge-intensive tasks," in *NeurIPS*, 2023.

[9] S. Piao and S. Park, "Tinythinker: Distilling reasoning through coarse-to-fine knowledge internalization with self-reflection," in *NAACL (Long Papers)*. Association for Computational Linguistics, 2025, pp. 6069–6087.

[10] H. Song, H. Lee, J. Shin, S. Cho, C. Ko, and J. C. Park, "Does rationale quality matter? enhancing mental disorder detection via selective reasoning distillation," in *ACL (Findings)*, ser. Findings of ACL, vol. ACL 2025. Association for Computational Linguistics, 2025, pp. 21 738–21 756.

[11] OpenAI, "GPT-4 technical report," *CoRR*, vol. abs/2303.08774, 2023.

[12] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *CoRR*, vol. abs/1503.02531, 2015.

[13] Y. Gu, L. Dong, F. Wei, and M. Huang, "Minillm: Knowledge distillation of large language models," in *ICLR*. OpenReview.net, 2024.

[14] J. Ko, S. Kim, T. Chen, and S. Yun, "Distillm: Towards streamlined distillation for large language models," in *ICML*. OpenReview.net, 2024.

[15] C. Hsieh, C. Li, C. Yeh, H. Nakhost, Y. Fujii, A. Ratner, R. Krishna, C. Lee, and T. Pfister, "Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes," in *ACL (Findings)*. Association for Computational Linguistics, 2023, pp. 8003–8017.

[16] X. Li, S. He, F. Lei, J. JunYang, T. Su, K. Liu, and J. Zhao, "Teaching small language models to reason for knowledge-intensive multi-hop question answering," in *ACL (Findings)*. Association for Computational Linguistics, 2024, pp. 7804–7816.

[17] H. Liao, S. He, Y. Hao, X. Li, Y. Zhang, J. Zhao, and K. Liu, "Skintern: Internalizing symbolic knowledge for distilling better cot capabilities into small language models," in *COLING*. Association for Computational Linguistics, 2025, pp. 3203–3221.

[18] H. Liao, S. He, Y. Xu, Y. Zhang, K. Liu, and J. Zhao, "Neural-symbolic collaborative distillation: Advancing small language models for complex reasoning tasks," in *AAAI*. AAAI Press, 2025, pp. 24 567–24 575.

[19] T. Mihaylov, P. Clark, T. Khot, and A. Sabharwal, "Can a suit of armor conduct electricity? A new dataset for open book question answering," in *EMNLP*. Association for Computational Linguistics, 2018, pp. 2381–2391.

[20] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord, "Think you have solved question answering? try arc, the AI2 reasoning challenge," *CoRR*, vol. abs/1803.05457, 2018.

[21] Y. Bisk, R. Zellers, R. L. Bras, J. Gao, and Y. Choi, "PIQA: reasoning about physical commonsense in natural language," in *AAAI*. AAAI Press, 2020, pp. 7432–7439.

[22] B. Y. Lin, Z. Wu, Y. Yang, D. Lee, and X. Ren, "Riddlesense: Reasoning about riddle questions featuring linguistic creativity and commonsense knowledge," in *ACL/IJCNLP (Findings)*, ser. Findings of ACL, vol. ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 1504–1515.

[23] Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, and X. Lu, "Pubmedqa: A dataset for biomedical research question answering," in *EMNLP/IJCNLP (1)*. Association for Computational Linguistics, 2019, pp. 2567–2577.

[24] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos, Y. Almirantis, J. Pavlopoulos, N. Baskiotis, P. Gallinari, T. Artières, A. N. Ngomo, N. Heino, É. Gaussier, L. Barrio-Alvers, M. Schroeder, I. Androutsopoulos, and G. Paliouras, "An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition," *BMC Bioinform.*, vol. 16, pp. 138:1–138:28, 2015.

[25] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," in *ICLR*. OpenReview.net, 2022.

[26] P. Wang, A. Chan, F. Ilievski, M. Chen, and X. Ren, "PINTO: faithful language reasoning using prompt-generated rationales," in *ICLR*. OpenReview.net, 2023.