Re-TASK: Revisiting LLM Tasks from Capability, Skill, and Knowledge Perspectives

Anonymous ACL submission

Abstract

The Chain-of-Thought (CoT) paradigm has become a pivotal method for solving complex problems with large language models (LLMs). However, its application to domain-specific tasks remains challenging, as LLMs often fail to decompose tasks accurately or execute subtasks effectively. This paper introduces the **Re-TASK** framework, a novel theoretical model that **Re**visits LLM **T**asks from c**A**pability, **S**kill, and **K**nowledge perspectives, drawing on the principles of Bloom's Taxonomy and Knowledge Space Theory. While CoT provides a workflow-centric perspective on tasks, Re-TASK introduces a Chain-of-Learning (CoL) paradigm that highlights task dependencies on specific capability items, further broken down into their constituent knowledge and skill components. To address CoT failures, we propose a Re-TASK prompting strategy, which strengthens task-relevant capabilities through targeted knowledge injection and skill adaptation. Experiments across diverse domains demonstrate the effectiveness of Re-TASK. In particular, we achieve improvements of 45.00% on Yi-1.5-9B and 24.50% on Llama3-Chinese-8B for legal tasks. These results highlight the potential of Re-TASK to significantly enhance LLM performance and its applicability in specialized domains. We release our code and data at this url.

1 Introduction

As the scale of large language models (LLMs) continues to increase, their general capabilities in natural language processing (NLP) tasks have shown substantial improvements. However, despite these advances, these models often struggle with complex reasoning tasks, particularly those that are domain-specific. The Chain-of-Thought (CoT) technique (Wei et al., 2022; Wang et al., 2023b; Zhou et al., 2023) has emerged as a promising paradigm by decomposing complex tasks into a series of subtasks in a divide-and-conquer manner. Yet, the application of CoT to domain-specific tasks faces significant challenges in both task decomposition (Kambhampati, 2024) and subtask execution (Lightman et al., 2024) due to a lack of domain knowledge and specialized capabilities.

The idea that an individual's capabilities directly influence task performance is well supported by educational theories, notably Bloom's Taxonomy (Bloom, 2010) and Knowledge Space Theory (KST) (Doignon and Falmagne, 1985). Bloom's Taxonomy outlines how educational objectives are achieved through structured instructional activities, each involving specific knowledge and cognitive processes. Similarly, KST emphasizes the sequential dependencies between learning items, forming *"learning pathways"* that guide learners from foundational knowledge to mastery.

Building on these insights into the Chain-of-Learning (CoL), we introduce the Re-TASK framework, which revisits LLM tasks through the lenses of capability, skill, and knowledge (see Figure 1(b)). This framework posits that the successful completion of tasks[1] depends on sequentially mastering multiple capability items, with each item further dissected into its constituent aspects of knowledge and skills.

Our framework argues that failures in the CoT paradigm, particularly in task decomposition and subtask execution, stem from a lack of the necessary capabilities, due to either insufficient knowledge or inadequate knowledge-skill adaptation (abbreviated as skill adaptation). First, LLMs may lack relevant knowledge due to limited access to proprietary data or issues related to data timeliness. Second, even when knowledge is available, LLMs often struggle to apply it effectively because of inadequate knowledge-skill adaptation, resulting in

[1] We use the terms "task" and "subtask" interchangeably, depending on the context.

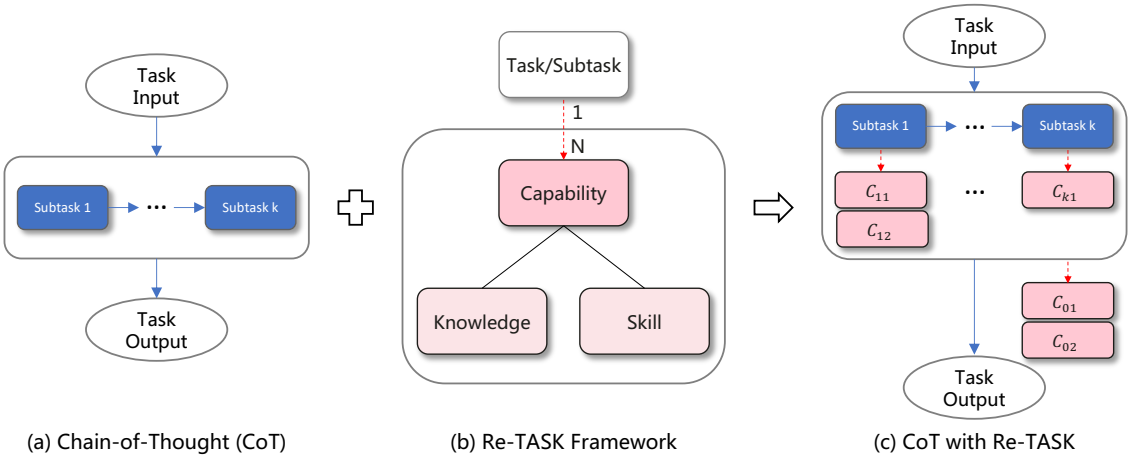

Figure 1: The Chain-of-Thought (CoT) provides a workflow perspective on tasks (blue arrow), while the Re-TASK framework introduces a Chain-of-Learning view (red dashed arrow), demonstrating how tasks and subtasks depend on various N capability items. Combining CoT with Re-TASK enhances CoT's performance in both task decomposition and subtask execution. Here, C_{ij} represents the capability item associated with subtask i (where $i = 1, \cdots, k$), and C_{0j} is associated with the overall task for task decomposition.

suboptimal performance. While techniques such as retrieval-augmented generation (Lewis et al., 2020) can inject knowledge into the context, models may still fail if there is insufficient skill adaptation to effectively utilize the retrieved knowledge.

To address these issues, we propose integrating the CoT paradigm with the Re-TASK framework to improve LLM performance, as shown in Figure 1. Specifically, we identify core capability items linked to the overall task and its corresponding subtasks, then strengthen these capabilities through targeted knowledge injection and skill adaptation using a deliberately designed prompting strategy, Re-TASK prompting (see Figure 2). The capability items represent demonstrations of knowledge-skill adaptation, such as conceptual knowledge understanding and procedural knowledge applying. Notably, knowledge itself can be treated as a special capability item, as in the case of knowledge injection through knowledge recalling or retrieving. We then adopt in-context learning (ICL) techniques (Brown et al., 2020; Dong et al., 2022) to enhance the corresponding capabilities by carefully arranging these demonstrations within the prompt.

We conduct extensive experiments with open-source LLMs to evaluate the effectiveness of the Re-TASK framework in enhancing CoT performance across various domain-specific tasks. By incorporating relevant capability items that inject domain knowledge or improve skill adaptation, we observe significant improvements in task performance. Additionally, we extend our experiments to LLMs of different scales, demonstrating that while model capabilities generally improve with scale, the Re-TASK framework continues to enhance performance effectively.

Our main contributions are summarized as follows:

- We introduce the Re-TASK framework, a novel theoretical model that revisits LLM tasks from the perspectives of capability, skill, and knowledge, offering a Chain-of-Learning view of tasks.
- Our research reveals that many failures of the CoT approach in addressing domain-specific tasks stem from insufficient knowledge or inadequate skill adaptation.
- We propose the Re-TASK prompting strategy, which integrates CoT with the Re-TASK framework to enhance LLM performance using in-context learning techniques.
- Extensive experimental results demonstrate the effectiveness of the Re-TASK framework.

2 Related Work

2.1 Educational Theories

Bloom's Taxonomy (Bloom, 2010) provides a foundational framework that links learning objectives with instructional activities. It posits that achieving learning goals requires the completion of multiple interconnected activities, categorized along two key

dimensions: knowledge and cognitive processes. The knowledge dimension outlines four types of knowledge: factual, conceptual, procedural, and metacognitive. The cognitive process dimension establishes a hierarchy of cognitive skills, encompassing six levels—remember, understand, apply, analyze, evaluate, and create—each linked to specific cognitive processes, totaling 19 distinct actions. By integrating these dimensions, Bloom's Taxonomy serves as a comprehensive guide for educators to design curricula and instructional strategies that foster deeper understanding and encourage higher-order thinking in students.

Knowledge Space Theory (KST) (Doignon and Falmagne, 1985; Falmagne et al., 2013; Cosyn et al., 2021) offers a mathematical framework for modeling and assessing learners' knowledge within a specific domain. It identifies various knowledge states, defined as sets of problems or concepts that a learner can successfully solve or understand. The entirety of these possible knowledge states forms the knowledge structure, which delineates the relationships among different states. Learning pathways are the routes learners can take to transition from one knowledge state to another. By utilizing KST, educators can design effective educational interventions and personalized learning experiences, optimizing the learning process for each individual learner.

Building on these foundational educational theories, we propose the Re-TASK framework, which elucidates the dependence of tasks on various capability items. Each capability item is further broken down into its constituent aspects of knowledge and skills, highlighting the intricate relationships that contribute to task performance.

2.2 Knowledge and Skills in LLMs

Several studies have explored LLMs from the perspectives of knowledge and skill. KoLA (Yu et al., 2023) emphasizes the importance of world knowledge for LLMs and establishes a knowledge-oriented evaluation benchmark. In its approach to ability modeling, KoLA simplifies and selects from Bloom's learning theories to form four levels of knowledge-capability assessment: knowledge memorization, knowledge understanding, knowledge applying, and knowledge creating. Skill-it (Chen et al., 2024) posits that language models naturally acquire a sequence of skills from training data and formalizes the notion of a skill and an ordered set of skills in terms of associated data, differentiating this approach from traditional curriculum learning (Bengio et al., 2009), which focuses on training models using progressively difficult examples. In continual pre-training experiments, Skill-it's ordered learning of skills achieves faster convergence of validation loss compared to random sampling. RA-DIT (Lin et al., 2023) introduces a lightweight fine-tuning methodology that improves retrieval-augmented language models by enhancing both the relevance of retrieved knowledge and its effective utilization, marking a specialized form of knowledge and skill enhancement. MMLU (Hendrycks et al., 2020) serves as a benchmark designed to measure the possession of world knowledge and problem-solving abilities. Reflecting on these developments, Bengio and Hu (2023) have emphasized the integration of the world model and the inference machine in current LLMs. They suggest that to reason effectively, a robust world knowledge model and a powerful inference machine are necessary, advocating for their separation and simultaneous development to enhance reasoning capabilities.

2.3 Prompting Strategies

Various prompting strategies have been proposed to enhance model performance in solving complex, domain-specific tasks. One prominent approach is Chain-of-Thought (CoT) (Wei et al., 2022; Wang et al., 2023b; Zhou et al., 2023), which decomposes a complex task into simpler subtasks, utilizing a divide-and-conquer strategy. Another notable approach involves integrating LLMs with knowledge. Some methods achieve this by retrieving knowledge from external sources, such as Retrieval-Augmented Generation (RAG) (Lewis et al., 2020; Fan et al., 2024), while others enable the models to generate knowledge internally, rather than relying on retrieval (Liu et al., 2021). In-Context Learning (ICL) techniques (Brown et al., 2020; Dong et al., 2022) represent a significant advancement in prompting; ICL utilizes examples within the prompt itself, enabling the model to learn from context without the need for explicit retraining or fine-tuning.

In contrast, we propose Re-TASK prompting, which integrates CoT prompting with the Re-TASK framework. This approach leverages ICL techniques to enhance corresponding capabilities by carefully arranging demonstrations of capability items within the prompt, ultimately leading to improved overall task performance.

3 Re-TASK

3.1 Re-TASK Framework

We begin by defining several key concepts within our framework: tasks, capability items, knowledge, and skills. We then elucidate how these elements interconnect to establish the structured Re-TASK framework, as illustrated in Figure 1(b). These concepts parallel Bloom's Taxonomy, where educational objectives—comparable to tasks in our framework—are systematically achieved through structured instructional activities, analogous to capability items. Each instructional activity involves the acquisition of specific types of knowledge and engages distinct cognitive processes, thereby facilitating knowledge-skill adaptation. The successful completion of an educational objective, or task, depends on mastering various capability items, each developed through these instructional activities. Knowledge Space Theory (KST) further reinforces this structured approach by highlighting sequential dependencies among capability items.

Definition 1 ***(Task)*** *A task* $\mathbf{T}$ *is defined as a specific objective that LLMs are designed to achieve, characterized by a mapping from input* x *to output* y*, facilitated by a task instruction* I *and an optional context* ctx*. Formally, this relationship is expressed as* $\mathbf{T}(ctx; I; x) = y$*, where the semicolon denotes concatenation of inputs.*

The optional context ctx can be leveraged for knowledge injection or skill adaptation. By identifying corresponding capability items and incorporating a list of capability item demonstrations into ctx, this method aligns closely with in-context learning. In the Chain-of-Thought (CoT) paradigm, a task can be decomposed into a series of subtasks.

Definition 2 ***(Knowledge)*** *A knowledge point* $\mathbf{K}$ *is defined as a text segment containing domain-specific knowledge that is essential for the performance of a task/subtask. In the context of LLMs, the knowledge* $\mathbf{K}$ *can also be implicit knowledge encoded within the model's parameters.*

According to Bloom's Taxonomy, we consider three types of domain knowledge: factual, conceptual, and procedural[2]. Each type plays a distinct role in task execution, contributing differently to the LLM's ability to process and respond to task-specific demands.

[2]Metacognitive knowledge is beyond the scope of LLMs and is not our primary focus.

Definition 3 ***(Skill)*** *A skill* $\mathbf{S}$ *corresponds to the cognitive processes in Bloom's Taxonomy and is developed through related instructional activities, including knowledge recalling/retrieving, understanding, applying, and others.*

Definition 4 ***(Capability Item)*** *A capability item* $\mathbf{C}$*, corresponding to the concept of instructional activities in Bloom's Taxonomy, is a specific exercise or demonstration designed to guide LLMs in applying a particular skill* $\mathbf{S}$ *to the relevant knowledge* $\mathbf{K}$*, thereby facilitating knowledge-skill adaptation.*

Successfully completing a task $\mathbf{T}$ requires the sequential mastery of multiple capability items. These items can be conceptualized as a chain of learning, where dependencies among them are clearly defined. Figure 1(c) illustrates these dependencies. A task generally involves overall procedural knowledge (C_{01}), with its resolution corresponding to a capability item (C_{02}) that applies this knowledge in a manner akin to a CoT process. This procedural knowledge is further segmented into subtasks, each linked to specific knowledge and involving different capability items C_{ij}.

Note that the knowledge $\mathbf{K}$ can be treated as a special capability item with the default skill of knowledge recalling or retrieving. Consequently, the task and its subtasks depend on two types of capability items: knowledge recalling (the knowledge itself) and knowledge-skill adaptation (e.g., knowledge understanding and applying). For example, a sentencing prediction task in law can be decomposed into two subtasks: the first involves assessing the severity of the victim's injuries, while the second determines the sentence duration based on this severity. For the first subtask, the capability item C_{11} could represent knowledge of guidelines for assessing severity (*knowledge recalling*), while C_{12} could demonstrate the application of C_{11} to assess the severity of the victim's injuries (*knowledge-skill adaptation*), formatted as *{Instruction; Input: case description; Output: severity}*.

3.2 Capability Item Construction

To effectively improve the performance of LLMs in domain-specific tasks, identifying key capability items is crucial. Our primary focus is on retrieving relevant knowledge and enhancing the understanding of conceptual knowledge along with the applying of procedural knowledge (i.e., knowledge-skill adaptation). Below, we present exemplary capability items for each category:

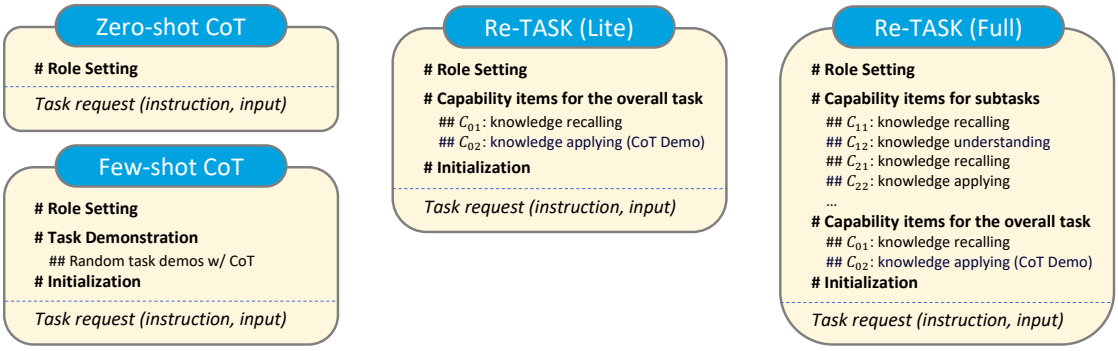

Figure 2: Comparison of prompting strategies: Zero-shot CoT, Few-shot CoT, Re-TASK (Lite) prompting with only capability items for the overall task, and Re-TASK (Full) prompting incorporating all available capability items. In Re-TASK prompting, a task or subtask may be associated with any number of capability items. Note that other strategies, such as Self-Consistency (SC), are excluded for brevity.

1) Knowledge retrieval: This involves identifying the relevant knowledge points for a given task or subtask and retrieving them from external sources. It may also include recalling internal knowledge points stored within the LLM.

2) Instances of conceptual knowledge: This involves providing examples that illustrate conceptual knowledge, helping to deepen and reinforce understanding of conceptual knowledge.

3) Execution of procedural knowledge: This capability is crucial for tasks that involve following ordered steps or procedures, such as technical troubleshooting, recipe preparation, or complex calculations. It represents the practical application of procedural knowledge.

Notably, identifying the capability items associated with a task is a straightforward process. For each task, we begin by determining the relevant knowledge points, followed by identifying the corresponding skills required for effective task resolution, either understanding or applying.

3.3 Re-TASK Prompting

By combining CoT with the Re-TASK framework, we can identify core capability items linked to the overall task and its corresponding subtasks, subsequently strengthening these capabilities to enhance subtask performance and, ultimately, overall task performance. Specifically, we enhance the capability items through targeted knowledge injection and skill adaptation using a deliberately designed prompting strategy known as Re-TASK prompting, as illustrated in Figure 2. The capability items serve as demonstrations of knowledge-skill adaptation, including knowledge recalling/retrieving, knowledge understanding, and knowledge applying.

We carefully arrange the demonstrations within the prompt according to their dependencies, following the chain of learning. Specifically, we sequence the capability items C_{ij} for each subtask i. If multiple items are associated with the same subtask, we prioritize the knowledge itself (i.e., the capability item of knowledge retrieval) first, followed by more advanced items such as knowledge understanding and applying. Finally, we include the overall procedural knowledge C_{01} along with its applying C_{02} at the end of the prompt.

4 Experimental Setup

4.1 Tasks and Datasets

We selected the sentencing prediction task in the law domain, the financial course examination task in the finance domain, and multiple-choice question tasks in three STEM domains—mathematics, biology, and physics—to validate the effectiveness of the Re-TASK framework.

The sentencing prediction task in the law domain involves predicting the appropriate sentencing range for criminal offenders based on the factual descriptions in criminal cases. This process requires a comprehensive understanding of legal statutes and sentencing principles, making it a suitable challenge for validating the Re-TASK framework. For dataset construction, we utilized the publicly available CAIL dataset (China AI Law Challenge) (Xiao et al., 2018), sampling 200 test instances. More details are presented in the Appendix C.1.

In the finance domain, we selected the financial course examination task in the FinanceIQ dataset (Zhang et al., 2023). The FinanceIQ dataset assesses LLMs' knowledge and reasoning abilities in

financial contexts, evaluating their grasp of domain-specific knowledge. The financial dataset consists of multiple-choice questions, and the test set for the FinanceIQ dataset contains 178 instances.

To evaluate the reasoning capabilities in STEM fields, we conducted experiments on the MMLU-Mathematics, Biology, and Physics benchmarks (Hendrycks et al., 2020) to evaluate the corresponding reasoning capabilities. The test sets for these benchmarks contain 276, 144, and 102 multiple-choice questions, respectively. During dataset construction, a few mathematics questions entered an infinite loop while generating capability items. As a result, we removed these instances from the mathematics dataset. Further details can be found in the Appendix C.3.

4.2 Settings

Given that both the legal and financial datasets are in Chinese, we opted for popular Chinese LLMs. Specifically, we selected the chat versions of Qwen1.5 (Bai et al., 2023), Llama3-Chinese (Cui, 2023), and Yi-1.5 (Young et al., 2024) for validation. For the mathematics, biology, and physics datasets, which are in English, we employed a different set of LLMs, including Llama3 (Dubey et al., 2024), Mistral (Jiang et al., 2023), and Qwen1.5, to validate our results. To further verify the scalability of our framework, we conducted experiments with LLMs of varying scales, using the Qwen1.5 series with 7B, 14B, and 32B parameters on the legal dataset. More setting details are offered in the Appendix B.2.

We developed Re-TASK prompting strategies that integrate demonstrations of capability items to validate their effectiveness in improving the performance of CoT on domain-specific tasks. Our baselines include Zero-shot CoT, Few-shot CoT, Plan-and-Solve (Wang et al., 2023a) and STEP-BACK (Zheng et al., 2023). We evaluated the performance of Few-shot CoT under two settings: 1-shot and 3-shot learning, to align with the Re-TASK (Lite) and Re-TASK (Full) strategies, respectively. Additionally, we evaluated the performance of Zero-shot CoT with self-consistency. The prompt templates are illustrated in Figure 2. Some prompt templates are shown in the Appendix D.

4.3 Construction of Capability Items

We utilized LLMs to assist in decomposing tasks and generating the capability items involved in Re-TASK. First, we predefined the capability types for each task based on the definition in Section 3.2 and then leveraged larger LLMs, which are presumed to possess sufficient domain knowledge, to automatically generate these items.

We began by employing LLMs to decompose the tasks, which resulted in the identification of the overall procedural knowledge (C_{01}). Next, we instructed the LLMs to create a CoT demonstration using the generated knowledge as a knowledge applying capability item (C_{02}). Task decomposition also guides the creation of capability items for each subtask. For each subtask i, we instructed the LLMs to generate relevant conceptual or procedural knowledge (C_{i1}). For conceptual knowledge, we requested illustrative examples to enhance understanding, while for procedural knowledge, we sought CoT demonstrations to illustrate effective applying (C_{i2}).

For financial tasks, we followed the complete procedure to generate capability items for both the overall task and its corresponding subtasks. In contrast, for other tasks, where the subtasks are relatively straightforward and do not require complex knowledge, we only generated capability items for the overall tasks.

5 Experimental Results

We conducted extensive experiments across five domains to validate the effectiveness of Re-TASK.

5.1 Law Domain

The results of the sentencing prediction task are presented in Table 1. Notably, Re-TASK (Lite) achieves the best performance across all settings, surpassing Zero-shot CoT by an average of 27.17%. Re-TASK (Lite) also outperforms the self-consistency version of Zero-shot CoT by a substantial margin of 26.84% on average. Specifically, Re-TASK (Lite) demonstrates a remarkable improvement of 45.00% with the Yi-1.5-9B model. Furthermore, when the knowledge item (C_{01}) is excluded from the demonstration count, Re-TASK (Lite), with a single demonstration, shows significant gains compared to the 1-shot CoT strategy. Step-Back also achieves a notable improvement, outperforming Zero-shot CoT by 18.00%, though it still lags behind Re-TASK (Lite) by 9.17%. These results underscore the effectiveness of the Lite version of Re-TASK prompting in the legal domain.

We further compared the token lengths generated by different prompt strategies, including both the

Table 1: Comparison of accuracy (%) across Zero-shot CoT, Few-shot CoT, and Re-TASK strategies in the law domain. "Zero-shot CoT + SC" refers to Zero-shot CoT with self-consistency, and "n-shot CoT" refers to Few-shot CoT with n randomly selected demonstrations.

		Llama3-Chinese-8B	Yi-1.5-9B	Qwen1.5-7B	Average Gain
Traditional CoT	Zero-shot CoT	54.00	40.00	33.50	-
	Zero-shot CoT + SC	54.50	40.50	33.50	+0.33
	1-shot CoT	53.67	66.50	36.17	+9.61
	3-shot CoT	56.33	70.17	38.50	+12.50
	Plan-and-Solve	54.50	33.50	45.00	+1.83
	Step-Back	72.50	72.50	36.50	+18.00
Re-TASK	**Re-TASK (Lite)**	**78.50**	**85.00**	**45.50**	**+27.17**

Table 2: Comparison of token length acorss various prompting strategies in the law domain.

		Llama3-Chinese-8B	Yi-1.5-9B	Qwen1.5-7B
Traditional CoT	Zero-shot CoT	526	510	431
	1-shot CoT	1691	1185	1007
	3-shot CoT	3104	2292	2176
Re-TASK	**Re-TASK (Lite)**	**1291**	**1071**	**967**

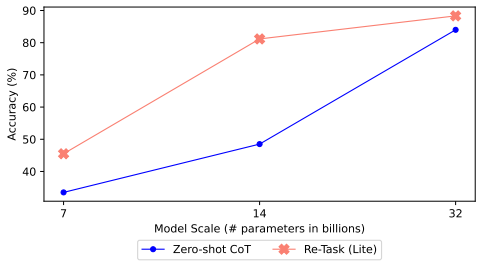

Figure 3: Performance comparison of zero-shot CoT and Re-TASK (Lite) across different scales of Qwen1.5 models (7B, 14B, and 32B) on the sentencing prediction task.

prompt and the completion, to assess efficiency. As shown in Table 2, the inclusion of demonstrations generally resulted in an increase in token length. Specifically, the token length of Re-TASK (Lite) was comparable to that of 1-shot CoT and shorter than that of 3-shot CoT.

5.1.1 Model Scaling

To investigate the impact of model scales, we conducted experiments on the Qwen1.5 series. The results, presented in Figure 3, reveal a consistent trend: as model size increases, the performance of both prompting strategies improves. Notably, the inclusion of capability items led to significant and consistent gains in task performance. These findings underscore the limitations of LLMs in handling domain-specific tasks and demonstrate that our Re-TASK framework delivers substantial benefits even as model size scales.

5.2 Financial Domain

As shown in Table 3, Re-TASK exhibits substantial performance gains across all models in the FinanceIQ task. Baseline methods without demonstrations, including Zero-shot CoT, Zero-shot CoT with self-consistency, Step-Back, and Plan-and-Solve strategies, show minimal improvement. In contrast, Re-TASK (Lite) and Re-TASK (Full) achieve notable average improvements of 5.06% and 14.61% over Zero-shot CoT, respectively. Specifically, Re-TASK (Lite) demonstrates an improvement of 19.67% with the Yi-1.5-9B model and 16.29% with the Llama3-Chinese-8B model. Furthermore, when comparing the number of demonstrations, Re-TASK (Lite) outperforms the 1-shot CoT strategy by 1.46%, while Re-TASK (Full) surpasses the 3-shot CoT strategy by 11.31%, highlighting the effectiveness of our added capability item components. Collectively, these results underscore the pivotal role and remarkable efficacy of Re-TASK in improving task performance.

5.3 STEM Domains

The results for MMLU-Math, MMLU-Biology and MMLU-Physics are presented in Table 4. Re-TASK (Lite) demonstrates significant performance gains across all three datasets, particularly in the biology domain, where it outperforms Zero-shot CoT by an average of 18.29%. In contrast, the self-consistency and plan-and-solve methods show inconsistent performance, with occasional improvements but no consistent gains overall. Moreover,

Table 3: Comparison of accuracy (%) across different prompt strategies on FinancelQ dataset.

		Llama3-Chinese-8B	Yi-1.5-9B	Qwen1.5-7B	Average Gain
Traditional CoT	Zero-shot CoT	36.52	53.93	43.82	-
	Zero-shot CoT + SC	34.27	61.80	46.63	+2.81
	1-shot CoT	34.69	64.33	46.07	+3.60
	3-shot CoT	34.27	63.82	46.07	+3.30
	Plan-and-Solve	30.34	66.29	41.01	+1.12
	Step-Back	30.90	66.85	44.38	+2.62
Re-TASK	**Re-TASK (Lite)**	**38.20**	**61.80**	**49.44**	**+5.06**
	Re-TASK (Full)	**52.81**	**73.60**	**51.69**	**+14.61**

Table 4: Comparison accuracy (%) across different prompt strategies on STEM datasets.

Domains	Prompting Strategies	Llama3-8B	Mistral-7B	Qwen1.5-7B	Average Gain
Math	Zero-shot CoT	40.58	24.28	36.96	-
	Zero-shot CoT+SC	48.19	24.64	41.67	+4.23
	1-shot CoT	49.42	23.41	36.52	+2.51
	Plan-and-Solve	40.58	30.43	23.91	-2.30
	Step-Back	45.65	34.42	19.93	-0.60
	Re-TASK (Lite)	**51.81**	**28.99**	**43.84**	**+7.61**
Physics	Zero-shot CoT	57.84	37.25	42.16	-
	Zero-shot CoT+SC	58.82	39.22	37.25	-0.65
	1-shot CoT	60.78	45.10	42.16	+3.59
	Plan-and-Solve	55.88	42.16	41.18	+0.65
	Step-Back	30.39	34.31	30.39	-14.05
	Re-TASK (Lite)	**60.78**	**44.12**	**50.98**	**+6.21**
Biology	Zero-shot CoT	76.39	57.64	59.72	-
	Zero-shot CoT+SC	78.47	60.42	62.50	+2.55
	1-shot CoT	79.86	68.75	60.42	+5.09
	Plan-and-Solve	73.61	55.56	61.11	-1.16
	Step-Back	43.75	50.69	53.47	-15.28
	Re-TASK (Lite)	**88.19**	**79.17**	**81.25**	**+18.29**

Step-Back performs worse across the three datasets, likely due to suboptimal principles generated by the smaller scale models themselves. When comparing with the same number of added demonstrations, Re-TASK (Lite) outperforms 1-shot CoT across all three datasets. Notably, Re-TASK (Lite) achieves a notable 13.20% improvement in the biology dataset, highlighting its strong and consistent performance in STEM tasks.

Compared to the results in the legal domain, the improvements in the STEM datasets are relatively modest. This may be attributed to the fact that the capability items automatically generated by LLMs, which, while effective, may not be fully optimized. Nonetheless, our methods show that leveraging knowledge and capability items generated from larger models can be used to augment the performance of smaller models. Since our primary objective is to demonstrate the potential of our approach rather than to optimize the identification of capability items, these results still underscore the value of our method.

6 Conclusion

In this paper, we introduced the Re-TASK framework, a novel approach that revisits LLM tasks through the lenses of capability, skill, and knowledge, aiming to address the limitations of the Chain-of-Thought (CoT) paradigm in complex, domain-specific tasks. By integrating Re-TASK with CoT, we systematically enhanced the performance of LLMs through targeted knowledge injection and skill adaptation using a structured prompting strategy, Re-TASK prompting. Our extensive experiments across the law, finance, and STEM domains demonstrated that the Re-TASK framework significantly improves task performance, achieving substantial gains over baseline models and confirming the framework's potential to enhance LLM capabilities across diverse domains. Our research opens new avenues for practitioners to deepen their understanding, evaluation, and improvement of LLMs.

Limitations

While this study presents promising results, several limitations should be considered when interpreting the findings.

First, we rely on powerful LLMs to directly construct capability items for each task instance without incorporating the retrieval process. In a more practical scenario, an offline repository of capability items could be constructed based on a domain knowledge graph to support knowledge retrieval, along with corresponding capability items for knowledge understanding or applying. Retrieval-based methods (e.g., RAG) could then be employed to match capability items with tasks. This is part of our future work.

Second, although we used LLMs to automatically generate capability items, we did not focus on optimizing the generation process. As our primary objective was to demonstrate the potential of our approach rather than to refine the identification of capability items, we leave capability item optimization for future research.

Third, we did not thoroughly analyze the performance differences across various domains. These differences may stem from the extent to which LLMs possess domain-specific knowledge and their proficiency in applying corresponding skills, both of which influence the observed improvement ratios. A deeper investigation into these aspects would be a valuable direction for LLM diagnostics.

Finally, our study is limited to open-source models due to budget constraints, leaving the exploration of proprietary models to future work.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yoshua Bengio and Edward J Hu. 2023. Scaling in the service of reasoning & model-based ml. Blog.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Benjamin Samuel Bloom. 2010. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Mayee Chen, Nicholas Roberts, Kush Bhatia, Jue Wang, Ce Zhang, Frederic Sala, and Christopher Ré. 2024. Skill-it! a data-driven skills framework for understanding and training language models. *Advances in Neural Information Processing Systems*, 36.

Eric Cosyn, Hasan Uzun, Christopher Doble, and Jeffrey Matayoshi. 2021. A practical perspective on knowledge space theory: Aleks and its data. *Journal of Mathematical Psychology*, 101:102512.

Yiming Cui. 2023. Chinese LLaMA-Alpaca 3. https://github.com/ymcui/Chinese-LLaMA-Alpaca-3.

Jean-Paul Doignon and Jean-Claude Falmagne. 1985. Spaces for the assessment of knowledge. *International journal of man-machine studies*, 23(2):175–196.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, and et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jean-Claude Falmagne, Dietrich Albert, Christopher Doble, David Eppstein, and Xiangen Hu. 2013. *Knowledge spaces: Applications in education*. Springer Science & Business Media.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 6491–6501, New York, NY, USA. Association for Computing Machinery.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Subbarao Kambhampati. 2024. Can large language models reason and plan? *Annals of the New York Academy of Sciences*, 1534(1):15–18.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Rich James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, et al. 2023. Ra-dit: Retrieval-augmented dual instruction tuning. *arXiv preprint arXiv:2310.01352*.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2021. Generated knowledge prompting for commonsense reasoning. *arXiv preprint arXiv:2110.08387*.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023a. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. *arXiv preprint arXiv:2305.04091*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, et al. 2018. Cail2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-Li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, et al. 2023. Kola: Carefully benchmarking world knowledge of large language models. *arXiv preprint arXiv:2306.09296*.

Xuanyu Zhang, Qing Yang, and Dongliang Xu. 2023. Xuanyuan 2.0: A large chinese financial chat model with hundreds of billions parameters. *Preprint*, arXiv:2305.12002.

Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H Chi, Quoc V Le, and Denny Zhou. 2023. Take a step back: Evoking reasoning via abstraction in large language models. *arXiv preprint arXiv:2310.06117*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

A All Experimental Results

A.1 Model Scaling

The accuracy results for zero-shot CoT and Re-TASK (Lite) across various scales of Qwen1.5 models on the sentencing prediction dataset are shown in Table 5, while the corresponding results on the FinancelQ dataset are presented in Table 6. Re-TASK demonstrates effectiveness across models of different sizes. A consistent trend emerges across different capability metrics: as the model scale increases, the performance of all prompting strategies improves accordingly. This demonstrates that Re-TASK is effective even on larger-scale models. .

A.2 Efficiency

To evaluate the efficiency of Re-TASK and eliminate the influence of computing platform performance on efficiency assessment, we adopt the sum of input and output tokens for LLMs as the metric of efficiency. Results for the financial and mathematical datasets are provided here. Table 7 and Table 8 present the average token lengths for different methods on the FinanceIQ and MMLU-Math datasets across all questions. It can be observed that the token counts for Re-TASK are comparable to those of the corresponding 1-shot and 3-shot CoT strategies. Thus, the efficiency of Re-TASK aligns closely with that of few-shot CoT strategy, demonstrating that Re-TASK introduces no additional overhead while achieving significantly higher accuracy.

A.3 Case Study

To better demonstrate the effectiveness of Re-TASK, we conduct a case study on the FinanceIQ and MMLU-Math datasets.

Figures 4, 5, and 6 show the results of the same input-output example using three prompt strategies—Zero-shot CoT, Re-TASK (Lite), and Re-TASK (Full) —on the FinanceIQ dataset with the LLaMA3-8B-Chinese model. From the case, we can see that in the Zero-shot CoT strategy, the model fails to produce the correct answer because it lacks knowledge of the problem-solving steps and relevant knowledge about the topic, leading to repeated attempts and ultimately an incorrect result. In Re-TASK (Lite) strategy, while the model knows the correct formula for solving the problem, it makes errors during the numerical calculations, resulting in an incorrect answer as well. In Re-TASK (Full) strategy, the model applies the correct formula and arrives at the correct answer. This indicates that adding different capability items supplements the model's knowledge and enhances its abilities in domain tasks, demonstrating the effectiveness of the Re-TASK framework.

Table 9 and 10 show one pair of comparison of Zero-shot CoT and Re-TASK (Lite) strategies with Llama3-8B on MMLU-Math dataset. For the Zero-shot CoT, it can be seen that though the LLM can stimulate the knowledge of the Pythagorean theorem, the following reasoning and calculation is still wrong. For the Re-TASK, we incorporate knowledge of the Pythagorean theorem through the prompt and demonstrated the applying of this knowledge through one capability item, and the correct reasoning and calculation result can be obtained. This comparison intuitively represents that capability items can effectively improve the reasoning.

B Experimental Settings

B.1 Models

Seven models are utilized in this work: Qwen1.5-7B-Chat, Qwen1.5-14B-Chat, Qwen1.5-32B-Chat (Bai et al., 2023), Yi-1.5-9B-Chat (Young et al., 2024), LLaMA-Chinese-8B-Instruct (Cui, 2023), Llama3-8B (Dubey et al., 2024), and Mistral-7B (Jiang et al., 2023). The Qwen1.5, Yi-1.5, Llama3, and Mistral-7B models are obtained as official chat versions from Hugging Face, whereas the LLaMA-Chinese-8B-Instruct model is sourced from Modelscope. In this study, Qwen2.5-72B is used for task decomposition and skill item generation on Chinese datasets (FinanceIQ and sentencing prediction tasks), while Llama3.1-70B is employed for task decomposition and skill item generation on the English dataset (MMLU). The model sources and licenses are shown in Table 11.

B.2 Experimental Details

For the baseline prompts used in the experiments, we
shot CoT templates, with the Step_Back template and Plan-and-Solve template following the design from the original paper. The parameters for self-consistency were also based on the original work, and the prompt template used was consistent with the zero-shot CoT. Additionally, the selection of the normal one-shot and three-shot demonstrations was randomly chosen from the demonstration corpus.

Table 5: Accuracy comparison (%) across zero-shot CoT and Re-TASK (Lite) using various scales of Qwen1.5 models on the sentencing prediction dataset.

	Qwen1.5-7B	**Qwen1.5-14B**	**Qwen1.5-32B**
Zero-shot CoT	33.50	48.50	84.00
Re-TASK (Lite)	**45.5**	**81.17**	**88.33**

Table 6: Accuracy comparison (%) across different prompt strategies using various scales of Qwen1.5 models on FinancelQ dataset.

	Qwen1.5-7B	**Qwen1.5-14B**	**Qwen1.5-32B**
Zero-shot CoT	43.82	49.19	60.67
Re-TASK (Lite)	**49.44**	**57.30**	**61.80**
Re-TASK (Full)	**51.69**	**57.87**	**71.91**

For all experiments, except those involving Self-Consistency, the temperature is set to 0, and the top_p parameter is set to 1. Because the temperature is set to 0, the results of multiple runs are consistent. For Self-Consistency, following the settings from the original paper (Wang et al., 2022), the temperature is adjusted to 0.5, and top_p is set to 0.5. Self-consistency involves the results of voting after running 20 rounds of zero-shot CoT.

We utilize vllm (Kwon et al., 2023) as the inference framework for all deployments. For models under 70 billion parameters, we deploy each on a single 32GB AI graphics card. For models with 70 billion parameters or more, we utilize two 80GB AI graphics cards per model.

C Experimental Datasets

We used three publicly available datasets across five domains. The sources of the datasets used in this paper are shown in Table 12. The questions in these five datasets all require corresponding knowledge and skill adaptation to answer correctly, which is very suitable for verifying Re-TASK.

C.1 Details of Sentencing Prediction Datasets

We construct the sentencing prediction dataset using data from the Cail2018 competition (Xiao et al., 2018), which is sourced from publicly available criminal legal documents on "China Judgments Online". Each record in the dataset includes case descriptions and factual details extracted from the legal documents. Additionally, each case provides the applicable legal articles, the charges against the defendant, and the length of the sentencing. The dataset contains approximately 2.68 million criminal law documents, covering 183 different charges and 202 legal articles, with sentences ranging from 0 to 25 years, life imprisonment, and the death penalty. Our focus is solely on the task of sentencing prediction.

The sentencing prediction task involves estimating the length of a defendant's sentence based on the descriptions and facts provided in the criminal legal documents. This task integrates five key elements from the dataset: the facts of the crime, the charges, the referenced legal articles, the defendant's name, and the length of the sentence.

Due to the large size and complexity of the original CAIL2018 dataset, we performed some preprocessing to better validate the retask. We implemented the following processing steps:

1) Several commonly encountered criminal law articles were selected to serve as the knowledge base. The original dataset was then filtered to include only criminal cases relevant to these articles. Additionally, we excluded data instances where the sentence prediction months were inherently implied based on the sentence range.

2) A suitable task instruction was designed, and the output format of the task was standardized.

3) The specific month lengths of sentences were converted into three broader sentencing categories: A (under 3 years), B (3 to 10 years), and C (over 10 years).

4) A series of robust and effective test prompt templates were designed for the task.

From the CAIL 2018 dataset, we constructed a 200-instance training set and a 200-instance test set. The training set and test set are independent, with no overlapping instances. The distribution of the three sentencing categories (A, B, C) in both the training and test sets approximates a 1:1:1 ratio.

Table 7: Comparison of token length across various prompting strategies on FinanceIQ dataset

	Llama3-Chinese-8B	Yi-9B	Qwen1.5-7B
Zero-shot CoT	861	541	417
1-shot CoT	1138	844	767
3-shot CoT	1730	1478	1421
Re-TASK (Lite)	**1032**	**742**	**602**
Re-TASK (Full)	**1713**	**1707**	**1581**

Table 8: Comparison of token length across various prompting strategies on MMLU-Math dataset.

	Llama3-8B	Qwen1.5-7B	Mistral-7B
Zero-shot CoT	301	449	544
1-shot CoT	615	615	768
Re-TASK(Lite)	**628**	**673**	**769**

Examples of sentence prediction task is illustrated in Figure 7.

We use Qwen2.5-72B to assist in the generation of capability items for the sentencing prediction task. Specifically, for a given instance, including the question, options, and answer, we first use LLMs to generate the knowledge behind the question. Then, based on the generated knowledge, we create the capability item. If the knowledge is conceptual, we generate an explanation or a demonstration example of the knowledge. If the knowledge is procedural, we generate a demonstration of how the knowledge is applied. More prompt templates are shown in the Appendix D.

C.2 Details of MMLU datasets

We use MMLU-high-school Mathematics, MMLU-college Biology and MMLU-college Physics datasets in this paper. For the MMLU-Math dataset, we utilized the original 304 questions. We used Llama3.1-70B to generate knowledge, but for some questions, the model entered into a feedback loop, rendering the generated knowledge invalid. In some cases, repeated attempts still resulted in the same feedback loop. As a result, we removed these problematic question-answer pairs, ultimately filtering out 276 questions. For the MMLU-Physics and MMLU-Biology datasets, such issues were rare, so we kept all instances and used the original datasets.

We use Llama3.1-70B to assist in the generation of capability items for the MMLU datasets. The generation method is consistent with the one used for the sentencing prediction task in Appendix C.1.

C.3 Details of FinanceIQ datasets

For the legal and MMLU datasets, since the tasks are relatively simple, we only constructed capability items related to the task, without building capability items for subtasks. When conducting processing for the FinanceIQ dataset, we needed to build both task-related capability items and subtask-related capability items, which made the process more complex. However, our analysis revealed that a significant portion of the FinanceIQ dataset consists of knowledge-intensive questions that can be addressed solely by providing factual knowledge, without requiring high levels of comprehension or reasoning abilities. These questions can be directly answered using RAG (retrieval-augmented generation), making it difficult to identify specific subtasks or construct subtask items. Consequently, we initially employed the Qwen2.5-72B model to assist in a filtering process, which effectively eliminated simpler questions that could be answered directly with factual knowledge.

During the generation process, we encountered special situations: Some questions resulted in infinite answer loops, particularly with the Llama series models, preventing them from generating knowledge or capability items properly. This issue could affect the subsequent validation of the Re-TASK framework. Due to the issues, we removed this portion of the data from the FinanceIQ dataset. Ultimately, we constructed our financial dataset.

We utilize Qwen2.5-72B to assist in the task decomposition of instances in the FinanceIQ dataset. In our experiments, we found that the subtasks in FinanceIQ still present a certain level of difficulty, requiring subtask-level knowledge and capability items. Therefore, for the financial dataset,

Prompt and Response for Standard Zero-shot CoT on FinanceIQ Dataset.

 # Role: 你是一名经济领域的专家，请完成指定的任务。
(**# Role:** You are an expert in the field of economics, please complete the tasks specified by the user.)

任务描述：现在根据下列输入的问题，从4个选项中选择一个正确的选项，请将答案填在[选项]与[选项]之间，例如[选项]C[选项]，必须给出一个选项。
(**Task description:** Now, based on the following input question, select the correct option from the 4 options. Please fill in the answer between [options] and [options], for example [option] C [option], one option must be given.)

问题 (Question):
以5亿作为基础货币，当活期存款准备金率是10%，定期存款准备金率是2%，定期存款比率达到30%，超额准备金率为5%，通货比率为20%时，你知道货币存量（以亿为单位）应该是多少吗？
(With 500 million as the base currency, when the demand deposit reserve ratio is 10%, the fixed deposit reserve ratio is 2%,the fixed deposit ratio is 30%, the excess reserve ratio is 5%, and the currency ratio is 20%, Do you know what the currency stock (in billions) should be?)
选项 (Options): A:16.20, B:16.85, C:12.45, D:9.23

答案 (Answer):

首先，我们需要计算货币存量的基础值。货币存量的基础值=基础货币/(1+活期存款准备金率+定期存款准备金率+定期存款比率+超额准备金率+通货比率)。货币存量的基础值=5亿/(1+0.1+0.02+0.3+0.05+0.2)=5亿/1.67=3亿。因此，货币存量的基础值为3亿。

接下来，我们需要计算货币存量的超额准备金率。货币存量的超额准备金率=超额准备金率/(1+活期存款准备金率+定期存款准备金率+定期存款比率+通货比率)。货币存量的超额准备金率=0.05/(1+0.1+0.02+0.3+0.2)=0.05/1.67=3%。因此，货币存量的超额准备金率为3%。最后，我们需要计算货币存量的货币存量。货币存量=基础货币+超额准备金率*货币存量的超额准备金率。货币存量=5亿+0.05*3亿=5亿+1500万=5.15亿因此，货币存量为5.15亿。

(First, we need to calculate the base value of the money stock. Base value of the currency stock = Base currency/ (1 + demanddeposit reserve ratio + fixed deposit reserve ratio + fixed deposit reserve ratio + excess reserve ratio + currency ratio). The base value of the money stock = /(1+0.1+0.02+0.3+0.05+0.2)=5 million / 167 million = 300 million. Therefore, the base value of the money stock is 300 million

Next, we need to calculate the excess reserve ratio of the money stock. Excess Reserve Ratio of Money Stock = Excess ReserveRatio/ (1 + Demand Deposit Reserve Ratio + Fixed Deposit Reserve Ratio + Fixed Deposit Reserve Ratio + Currency Ratio). The excess reserve ratio of the money stock is =0.05/(1+0.1+0.02+0.3+0.2)=0.05/1.67=3% . As a result, the excess reserve ratio for the money stock is 3%.

Finally, we need to calculate the money stock of the money stock. Currency stock = Base currency + Excess reserve ratio * Excess reserve ratio of currency stock. Money stock = 500 million + 0.05 * 300 million = 500 million + 15 million = 515 million. Therefore, the money stock is 515 million.)

Figure 4: An example of the Prompt and the Response for Zero-shot CoT strategy on FinanceIQ Dataset.

we generated both task-related capability items and subtask-related capability items. We continue to use Qwen2.5-72B to help generate these capability items for FinanceIQ datasets. The generation process for task-related and subtask-related capability items is similar: we first provide the question, options, and answer to the LLM, allowing it to generate the knowledge; then, we feed the generated knowledge back to the model to generate the corresponding capability items.

D Experimental Prompts

D.1 Law Domain

The prompt templates using zero-shot CoT strategy, few-shot CoT strategy and Re-TASK (Lite) strategy for the sentencing prediction task in the law domain are shown in Figure 8, Figure 9, Figure 10.

D.2 STEM Domain

The prompt templates using zero-shot CoT strategy, few-shot CoT strategy and Re-TASK (Lite) strategy for the MMLU datasets in the STEM domains are shown in Table 13, Table 14 and Table 15.

The prompt templates for generating knowledge and capability items for the MMLU datasets in the STEM domains are shown in Table 16 and Table 17.

D.3 Financial Domain

The prompt templates using zero-shot CoT strategy, few-shot CoT strategy, Re-TASK (Lite) strategy and Re-TASK (Full) strategy in the financial domain are shown in Figure 11, Figure 12, Figure 13 and Figure 14.

The prompt templates for generating knowledge and capability items in the financial domains are shown in Figure 15 and Figure 16.

Prompt and Response for Re-Task (Lite) on FinanceIQ Dataset.

 # Role: 你是一名金融领域的专家，请通过对<Knowledge Application>的学习加深对响应的<Knowledge Recall>的应用能力，完成用户指定的任务。
(**# Role:** You are an expert in the field of economics, please deepen your application ability of<Knowledge Recall>through learning<Knowledge Application> and complete the tasks specified by the user.)

 # Capability Items for Overall Task:
Knowledge Recall
为了解决这道题，需要先理解题意，然后根据题意计算货币乘数，最后计算货币存量，可以得到答案。
(To solve this problem, it is necessary to first understand the meaning of the question, then calculate the money multiplier based on the meaning of the question, and finally calculate the money stock to obtain the answer.)

Knowledge Application
问题 (Question):
假设基础货币为4亿，活期存款准备金率为8%，定期存款准备金率为3%，定期存款比率为25%，超额准备金率为4%，通货比率为15%。根据这些条件，计算货币存量（以亿为单位）应该是多少？
(Assuming the base currency is 400 million, the reserve requirement ratio for current deposits is 8%, the reserve requirement ratio for fixed deposits is 3%, the fixed deposit ratio is 25%, the excess reserve requirement ratio is 4%, and the currency ratio is 15%. Based on these conditions, what should the currency stock (in billions) be calculated?)
选项 (Options): A:10.24, B:11.56, C:12.32, D:13.48
答案 (Answer): 首先，理解题意，题目要求计算在给定的基础货币、各种准备金率和通货比率条件下的货币存量。接着，计算货币乘数，货币乘数 = 1 / (活期存款准备金率 + 定期存款准备金率 * 定期存款比率 + 超额准备金率 + 通货比率) = 1 / (0.08 + 0.03 * 0.25 + 0.04 + 0.15) = 1 / (0.08 + 0.0075 + 0.04 + 0.15) = 1 / 0.2775 ≈ 3.604。最后，计算货币存量，货币存量 = 基础货币 * 货币乘数 = 4 * 3.604 ≈ 14.416。由于选项中没有14.416这个数值，最接近的是13.48，因此答案是[选项]D[选项]。但根据计算，最准确的答案应该是14.416，这表明选项中可能存在误差或需要重新评估选项设置。然而，基于提供的选项，最接近的答案是D。
(Firstly, understand the meaning of the question, which requires calculating the money stock under given conditions of base currency, various reserve ratios, and currency ratios. Next, calculate the currency multiplier, which is 1/(current deposit reserve ratio+fixed deposit reserve ratio * fixed deposit ratio+excess reserve ratio+currency ratio)=1/(0.08+0.03 * 0.25+0.04+0.15)=1/(0.08+0.0075+0.04+0.15)=1/0.2775 ≈ 3.604. Finally, calculate the money stock, where money stock=base currency * money multiplier=4 * 3.604 ≈ 14.416. Since there is no value of 14.416 in the options, the closest value is 13.48, so the answer is [option] D [option]. But according to calculations, the most accurate answer should be 14.416, indicating that there may be errors in the options or that the option settings need to be re evaluated. However, based on the provided options, the closest answer is D.)

 # Initialization
作为<Role>，你可以参考<Demonstration>完成用户给定的任务。
(As a<Role>, you can refer to<Demonstration>to complete the tasks given by the user.)

任务描述：现在根据下列输入的问题，从4个选项中选择一个正确的选项，请将答案填在[选项]与[选项]之间，例如[选项]C[选项]，必须给出一个选项。
(**Task description:** Now, based on the following input question, select the correct option from the 4 options. Please fill in the answer between [options] and [options], for example [option] C [option], one option must be given.)

问题 (Question):
以5亿作为基础货币，当活期存款准备金率是10%，定期存款准备金率是2%，定期存款比率达到30%，超额准备金率为5%，通货比率为20%时，你知道货币存量（以亿为单位）应该是多少吗？
(With 500 million as the base currency, when the demand deposit reserve ratio is 10%, the fixed deposit reserve ratio is 2%,the fixed deposit ratio is 30%, the excess reserve ratio is 5%, and the currency ratio is 20%, Do you know what the currency stock (in billions) should be?)
选项 (Options): A:16.20, B:16.85, C:12.45, D:9.23

答案 (Answer):
首先，我们需要计算货币存量的计算公式为：货币存量=基础货币/(1-活期存款准备金率-定期存款准备金率-超额准备金率)根据题目中的条件，我们可以计算货币存量：货币存量=5亿/(1-10%-2%-5%)=5亿/0.77=6.49亿所以，货币存量应该是6.49亿。答案：[选项]C[选项]
(First, we need to calculate the currency stock as follows: Currency stock = Base currency/ (1- Demand deposit reserve ratio - Time deposit reserve ratio - Excess reserve ratio) According to the conditions in the topic, we can calculate the money stock: Money stock = 500 million / (1- 10% - 2% - 5%) = 500 million / 0.77 = 649 million So, money stock should be 649 million. Therefore, the answer is C.)

Figure 5: An example of the Prompt and the Response for Re-TASK (Lite) strategy on FinanceIQ Dataset.

Prompt and Response for Re-Task (Full) on FinanceIQ Dataset.

 # Role: 你是一名金融领域的专家，请通过<Knowledge Example 1>的学习加深对响应的<Knowledge Recall 1>的理解能力，对通过对<Knowledge Application 2>的学习加深对响应的<Knowledge Recall 2>的应用能力，完成用户指定的任务。
(**# Role:** You are an expert in the field of economics, please deepen your understanding of <Knowledge Recall 1> through learning <Knowledge Example 1> and deepen your application ability of <Knowledge Recall> through learning <Knowledge Application> and complete the tasks specified by the user.)

 # Capability Items for Subtasks:
Knowledge Recall 1
基础货币、活期存款准备金率、定期存款准备金率、定期存款比率、超额准备金率、通货比率的概念及其对货币存量的影响。
(The concepts of base currency, reserve requirement ratio for current deposits, reserve requirement ratio for fixed deposits, fixed deposit ratio, excess reserve ratio, and currency ratio and their impact on the stock of money.)
Knowledge Example 1
[Please Put the Knowledge Example 1 of Subtasks Here.]

Knowledge Recall 2
货币存量 = 基础货币 / (活期存款准备金率 + 定期存款准备金率 * 定期存款比率 + 超额准备金率 + 通货比率)。
(Currency stock=base currency/(current reserve requirement ratio+fixed reserve requirement ratio * fixed deposit ratio+excess reserve requirement ratio+currency ratio).)
Knowledge Application 2
[Please Put the Knowledge Application 2 of the Subtasks Here.]

 # Capability Items for Overall Task:
Knowledge Recall
为了解决这道题，需要先理解题意，然后根据题意计算货币乘数，最后计算货币存量，可以得到答案。
(To solve this problem, it is necessary to first understand the meaning of the question, then calculate the money multiplier based on the meaning of the question, and finally calculate the money stock to obtain the answer.)
Knowledge Application
问题 (Question):
假设基础货币为4亿，活期存款准备金率为8%，定期存款准备金率为3%，定期存款比率为25%，超额准备金率为4%，通货比率为15%。根据这些条件，计算货币存量（以亿为单位）应该是多少？
(Assuming the base currency is 400 million, the reserve requirement ratio for current deposits is 8%, the reserve requirement ratio for fixed deposits is 3%, the fixed deposit ratio is 25%, the excess reserve requirement ratio is 4%, and the currency ratio is 15%. Based on these conditions, what should the currency stock (in billions) be calculated?)
选项 (Options): A:10.24, B:11.56, C:12.32, D:13.48
答案 (Answer): 首先，理解题意，题目要求计算在给定的基础货币、各种准备金率和通货比率条件下的货币存量。接着，计算货币乘数，货币乘数= 1 / (活期存款准备金率 + 定期存款准备金率 * 定期存款比率 + 超额准备金率 + 通货比率) = 1 / (0.08 + 0.03 * 0.25 + 0.04 + 0.15) = 1 / (0.08 + 0.0075 + 0.04 + 0.15) = 1 / 0.2775 ≈ 3.604。最后，计算货币存量，货币存量 = 基础货币 * 货币乘数 = 4 * 3.604 ≈ 14.416。由于选项中没有14.416这个数值，最接近的是13.48，因此答案是[选项]D[选项]。但根据计算，最准确的答案应该是14.416，这表明选项中可能存在误差或需要重新评估选项设置。然而，基于提供的选项，最接近的答案是D。
(Firstly, understand the meaning of the question, which requires calculating the money stock under given conditions of base currency, various reserve ratios, and currency ratios. Next, calculate the currency multiplier, which is 1/(current deposit reserve ratio+fixed deposit reserve ratio * fixed deposit ratio+excess reserve ratio+currency ratio)=1/(0.08+0.03 * 0.25+0.04+0.15)=1/(0.08+0.0075+0.04+0.15)=1/0.2775 ≈ 3.604. Finally, calculate the money stock, where money stock=base currency * money multiplier=4 * 3.604 ≈ 14.416. Since there is no value of 14.416 in the options, the closest value is 13.48, so the answer is [option] D [option]. But according to calculations, the most accurate answer should be 14.416, indicating that there may be errors in the options or that the option settings need to be re evaluated. However, based on the provided options, the closest answer is D.)

 # Initialization
作为<Role>，你可以参考<Demonstration>完成用户给定的任务。
(As a<Role>, you can refer to<Demonstration>to complete the tasks given by the user.)

任务描述：现在根据下列输入的问题，从4个选项中选择一个正确的选项，请将答案填在[选项]与[选项]之间，例如[选项]C[选项]，必须给出一个选项。
(**Task description:** Now, based on the following input question, select the correct option from the 4 options. Please fill in the answer between [options] and [options], for example [option] C [option], one option must be given.)

问题 (Question):
以5亿作为基础货币，当活期存款准备金率是10%，定期存款准备金率是2%，定期存款比率达到30%，超额准备金率为5%，通货比率为20%时，你知道货币存量（以亿为单位）应该是多少吗？
(With 500 million as the base currency, when the demand deposit reserve ratio is 10%, the fixed deposit reserve ratio is 2%,the fixed deposit ratio is 30%, the excess reserve ratio is 5%, and the currency ratio is 20%, Do you know what the currency stock (in billions) should be?)
选项 (Options): A:16.20, B:16.85, C:12.45, D:9.23

答案 (Answer):
首先根据货币乘数公式：货币乘数=1/(活期存款准备金率+定期存款准备金率*定期存款比率+超额准备金率+通货比率)。将题目中的数值代入公式计算：
1/(0.10+0.02*0.30+0.05+0.20)=1/(0.10+0.006+0.05+0.20)=1/0.286≈3.49 然后根据货币存量的计算公式：货币存量=基础货币*货币乘数。将题目中的数值代入公式得到答案是16.85亿元。因此，答案选B。
(First, according to the currency multiplier formula: currency multiplier = 1 / (Demand deposit reserve ratio + Fixed depositreserve ratio * Fixed deposit ratio + Excess reserve ratio + Currency ratio). Substitute the numerical value in the question into the formula: 1/(0.10+0.02*0.30+0.05+0.20)=1/(0.10+0.006+0.05+0.20)=1/0.286≈3.49 . Then, according to the calculation formula of currency stock: currency stock = base currency * currency multiplier. Substitute the value in the question into the formula and the answer is 1.685 billion yuan. Therefore, the answer is B.)

Figure 6: An example of the Prompt and the Response for Re-TASK (Full) strategy on FinanceIQ Dataset.

Table 9: An example of the Prompt and the Response for Zero-shot CoT strategy on MMLU-Math dataset.

Prompt and Response for Standard Zero-shot CoT
Input: # Role: You are an expert in the field of Math. Complete the task provided by the user. # Task Description: According to the following question and options, select one correct option. Please think step by step and follow the answer format, i.e., output the brief chain-of-thought beginning with 'Rationale: ' within 500 characters, and the final choice option beginning with 'Correct: ' from 'A' to 'D'. Question: The length of a rectangle is twice its width. Given the length of the diagonal is $5\sqrt5$, find the area of the rectangle. Options: A. 2500; B. 2; C. 50; D. 25 **Output:** Rationale: Let the width of the rectangle be x. Since the length is twice the width, the length is 2x. The diagonal is the hypotenuse of a right triangle with legs x and 2x. Using the Pythagorean theorem, we can write: x^2 + (2x)^2 = (5\sqrt5)^2 Simplifying the equation, we get: 5x^2 = 25 Dividing both sides by 5, we get: x^2 = 5 Taking the square root of both sides, we get: x = $\sqrt{5}$ Now, the area of the rectangle is: Area = length × width = 2x × x = 2($\sqrt{5}$)^2 = 2(5) = 10 Correct: D. 25

Table 10: An example of the Prompt and the Response for Re-TASK (Lite) strategy on MMLU-Math dataset.

Prompt and Response for Re-TASK Prompting
Input: # Role: You are an expert in the field of Math. Complete the task provided by the user. # Knowledge: The area of a rectangle is given by A = length * width, and the Pythagorean theorem states that for a right triangle, a^2 + b^2 = c^2, where c is the length of the hypotenuse (diagonal). # Demonstration: Question: In a rectangle, the length is 5 cm and the width is 12 cm. What is the length of the diagonal? Options: A. 13 cm; B. 14 cm; C. 15 cm; D. 16 cm Rationale: Step 1. We're given a rectangle with length 5 cm and width 12 cm. Step 2. We can use the Pythagorean theorem to find the length of the diagonal. Step 3. The diagonal is the hypotenuse of a right triangle with legs 5 cm and 12 cm. Step 4. Using the theorem, we get diagonal^2 = 5^2 + 12^2 = 25 + 144 = 169. Step 5. Taking the square root, we get diagonal = 13 cm. The correct answer is A. Correct: A # Task Description: According to the following question and options, refer to <# Knowledge> and <# Demonstration>, select one correct option. Please think step by step and follow the answer format in <# Demonstration>, i.e., output the brief chain-of-thought beginning with 'Rationale: ' within 500 characters, and the final choice option beginning with 'Correct: ' from 'A' to 'D'. Question: The length of a rectangle is twice its width. Given the length of the diagonal is $5\sqrt5$, find the area of the rectangle. Options: A. 2500; B. 2; C. 50; D. 25 **Output:** Rationale: Step 1. We're given a rectangle with length = 2 * width. Step 2. We're also given the length of the diagonal, which is $5\sqrt5$. Step 3. We can use the Pythagorean theorem to find the length and width. Step 4. Let the width be x, then the length is 2x. Step 5. Using the theorem, we get (2x)^2 + x^2 = ($5\sqrt5$)^2. Step 6. Simplifying, we get 4x^2 + x^2 = 125. Step 7. Combining like terms, we get 5x^2 = 125. Step 8. Dividing both sides by 5, we get x^2 = 25. Step 9. Taking the square root, we get x = 5. Step 10. Now we can find the area, which is length * width = 2x * x = 2 * 5 * 5 = 50. Correct: C

Table 11: Models, Sources and Licenses Used in This Work

Models	Model sources	License
Qwen1.5-7B	https://huggingface.co/Qwen/Qwen1.5-7B-Chat	Apache License 2.0
Qwen1.5-14B	https://huggingface.co/Qwen/Qwen1.5-14B-Chat	Apache License 2.0
Qwen1.5-32B	https://huggingface.co/Qwen/Qwen1.5-32B-Chat	Apache License 2.0
Yi-1.5-9B	https://huggingface.co/01-ai/Yi-1.5-9B-Chat	Apache License 2.0
Llama3-Chinese-8B	https://www.modelscope.cn/models/FlagAlpha/ Llama3-Chinese-8B-Instruct/summary	Apache License 2.0
Llama3-8B	https://www.modelscope.cn/models/FlagAlpha/ Llama3-8B-Instruct/summary	Apache License 2.0
Mistral-7B	https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2	Apache License 2.0
Llama3.1-70B	https://huggingface.co/meta-llama/Meta-Llama-3.1-70B-Instruct	llama3.1 license
Qwen2.5-72B	https://huggingface.co/Qwen/Qwen2.5-72B-Instruct	Qwen license

Table 12: Datasets and sources used in this work

Datasets	Sources
MMLU	https://huggingface.co/datasets/cais/mmlu
FinanceIQ	https://huggingface.co/datasets/Duxiaoman-DI/FinanceIQ
CAIL 2018	https://github.com/thunlp/CAIL?tab=readme-ov-file

Table 13: The prompt template of standard zero-shot CoT on MMLU-Math dataset.

Prompt template of Standard Zero-shot CoT
Role: You are an expert in the field of Math. Complete the task provided by the user. # Task Description: According to the following question and options, select one correct option. Please think step by step and follow the answer format, i.e., output the brief chain-of-thought beginning with 'Rationale: ' within 500 characters, and the final choice option beginning with 'Correct: ' from 'A' to 'D'. Question: [Please Put Your Question Here] Options: [Please Put Your Options Here]

Examples	
Instruction1	任务描述：根据下列输入，对被告人陈某进行刑期判决。 (Task description: According to the following input, sentence the defendant Chen.) 要求：从A,B,C三个刑期区间选项(A:三年以下有期徒刑、拘役或者管制；B:三年以上十年以下有期徒刑；C:十年以上有期徒刑、无期徒刑或死刑)中选出一个最合适的判决刑期区间，请将答案填在[刑期区间]与[eoa]之间，必须给出一个选项，请严格按要求的输出格式输出答案，不要输出任何无关内容或者解释。例如[刑期区间]C[eoa]。 (Requirement: Select the most appropriate sentence interval from the three sentence interval options A, B, and C (A: imprisonment of less than three years, detention or control; B: imprisonment of more than three years but less than ten years; C: imprisonment of more than ten years, life imprisonment or death penalty). Please fill in the answer between [SENTENCE] and [eoa]. You must give an option. Please output the answer strictly in the required output format. Do not output any irrelevant content or explanation. For example, [SENTENCE] C [eoa].)
Input1	淄博市张店区人民检察院指控，2013年10月21日8时许，被告人张某、张2某、陈某受栾某（已判刑）指使，对刘某实施殴打致刘某伤情构成轻伤二级。 (The People's Procuratorate of Zhangdian District, Zibo City, accused that at about 8:00 on October 21, 2013, the defendants Zhang, Zhang 2, and Chen were instructed by Luan (who has been sentenced) to beat Liu, causing Liu's injury to constitute a second-degree minor injury. The defendant Chen's behavior constitutes ××. The public prosecution agency has transferred relevant evidence to this court and requested this court to investigate the criminal responsibility of the defendant Chen in accordance with the provisions of ×××× of the Criminal Law of the People's Republic of China.)
Output1	[刑期区间]A[eoa] ([SENTENCE] A [eoa])
Instruction2	任务描述：根据下列输入，对被告人唐某进行刑期判决。 (Task description: According to the following input, sentence the defendant Tang.) 要求：从A,B,C三个刑期区间选项(A:三年以下有期徒刑、拘役或者管制；B:三年以上十年以下有期徒刑；C:十年以上有期徒刑、无期徒刑或死刑)中选出一个最合适的判决刑期区间，请将答案填在[刑期区间]与[eoa]之间，必须给出一个选项，请严格按要求的输出格式输出答案，不要输出任何无关内容或者解释。例如[刑期区间]C[eoa]。 (Requirement: Select the most appropriate sentence interval from the three sentence interval options A, B, and C (A: imprisonment of less than three years, detention or control; B: imprisonment of more than three years but less than ten years; C: imprisonment of more than ten years, life imprisonment or death penalty). Please fill in the answer between [SENTENCE] and [eoa]. You must give an option. Please output the answer strictly in the required output format. Do not output any irrelevant content or explanation. For example, [SENTENCE] C [eoa].)
Input2	2016年9月25日7时许，被告人唐某与黄某因公路堡坎问题发生口角，继而互相抓打。唐某手持锄头打向黄某，锄头打中正在劝架的被害人谭某后脑，致谭某当场昏迷。唐某拨打“110”电话报警，到案后如实供述了事发经过。经鉴定，被害人谭某损伤程度系重伤二级。 (At about 7:00 a.m. on September 25, 2016, the defendant Tang and Huang had an argument over a road embankment, and then they fought each other. Tang hit Huang with a hoe, and the hoe hit the back of the head of the victim Tan who was trying to stop the fight, causing Tan to faint on the spot. Tang called the "110" number to report the incident, and after being arrested, he truthfully confessed the incident. According to the appraisal, the victim Tan's injury was a second-degree severe injury.)
Output2	[刑期区间]B[eoa] ([SENTENCE] B [eoa])

Figure 7: Examples of the Sentence Prediction Task

Prompt Template for Zero-shot CoT in Law Domain.

 # Role: 你是一名法官，请完成用户指定的任务。
(**# Role:** You are a judge. Please complete the tasks specified by the user.)

任务描述：根据下列输入，对被告人李某某进行刑期判决。
要求：从A,B,C三个刑期区间选项(A:三年以下有期徒刑、拘役或者管制；B:三年以上十年以下有期徒刑；C:十年以上有期徒刑、无期徒刑或死刑)中选出一个最合适的判决刑期区间，请将答案填在[刑期区间]与[eoa]之间，必须给出一个选项，请严格按要求的输出格式输出答案，不要输出任何无关内容或者解释。例如[刑期区间]C[eoa]。
(**Task description:** According to the following input, sentence the defendant Li.
Requirements: Select the most appropriate sentence interval from the three sentence interval options A, B, and C (A: imprisonment of less than three years, detention or control; B: imprisonment of more than three years but less than ten years; C: imprisonment of morethan ten years, life imprisonment or death penalty). Please fill in the answer between [SENTENCE] and [eoa]. You must give an option. Pleaseoutput the answer strictly in the required output format. Do not output any irrelevant content or explanation. For example, [SENTENCE] C[eoa].)

输入 (Input): [Please Put Your Input Here.]

输出 (Output):

Figure 8: The Prompt Template for Zero-shot CoT in Law Domain.

Prompt Template for Few-shot CoT in Law Domain.

 # Role: 你是一名法官，请理解<参考法条>的基础上，完成用户指定的任务。
(**# Role:** You are a judge. Please understand the <reference law> and complete the tasks specified by the user.)

 # Demonstration:
[Please Put Your Demonstrations Here.]

任务描述：根据下列输入，对被告人李某某进行刑期判决。
要求：从A,B,C三个刑期区间选项(A:三年以下有期徒刑、拘役或者管制；B:三年以上十年以下有期徒刑；C:十年以上有期徒刑、无期徒刑或死刑)中选出一个最合适的判决刑期区间，请将答案填在[刑期区间]与[eoa]之间，必须给出一个选项，请严格按要求的输出格式输出答案，不要输出任何无关内容或者解释。例如[刑期区间]C[eoa]。
(**Task description:** According to the following input, sentence the defendant Li.
Requirements: Select the most appropriate sentence interval from the three sentence interval options A, B, and C (A: imprisonment of less than three years, detention or control; B: imprisonment of more than three years but less than ten years; C: imprisonment of morethan ten years, life imprisonment or death penalty). Please fill in the answer between [SENTENCE] and [eoa]. You must give an option. Pleaseoutput the answer strictly in the required output format. Do not output any irrelevant content or explanation. For example, [SENTENCE] C[eoa].)

输入 (Input): [Please Put Your Input Here.]

输出 (Output):

Figure 9: The Prompt Template for Few-shot CoT in Law Domain.

Prompt Template for Re-Task (Lite) in Law Domain.

 # Role: 你是一名法官，请完成用户指定的任务。
(**# Role:** You are a judge. Please complete the tasks specified by the user.)

 # 参考法条：
(**# Reference law**)
[Please Put the Knowledge Here.]

 # Demonstration:
[Please Put Your Capability Items for the Overall Task Here.]

任务描述：根据下列输入，对被告人李某某进行刑期判决。
要求：从A,B,C三个刑期区间选项(A:三年以下有期徒刑、拘役或者管制；B:三年以上十年以下有期徒刑；C:十年以上有期徒刑、无期徒刑或死刑)中选出一个最合适的判决刑期区间，请将答案填在[刑期区间]与[eoa]之间，必须给出一个选项，请严格按要求的输出格式输出答案，不要输出任何无关内容或者解释。例如[刑期区间]C[eoa]。
(**Task description:** According to the following input, sentence the defendant Li.
Requirements: Select the most appropriate sentence interval from the three sentence interval options A, B, and C (A: imprisonment of less than three years, detention or control; B: imprisonment of more than three years but less than ten years; C: imprisonment of morethan ten years, life imprisonment or death penalty). Please fill in the answer between [SENTENCE] and [eoa]. You must give an option. Pleaseoutput the answer strictly in the required output format. Do not output any irrelevant content or explanation. For example, [SENTENCE] C[eoa].)

输入 (Input): [Please Put Your Input Here.]

输出 (Output):

Figure 10: The Prompt Template for Re-TASK (Lite) in Law Domain.

Prompt Template for Zero-shot CoT on FinanceIQ Dataset.

 # Role: 你是一名经济领域的专家，请完成指定的任务。
(**# Role:** You are an expert in the field of economics, please complete the tasks specified by the user.)

任务描述：现在根据下列输入的问题，从4个选项中选择一个正确的选项，请将答案填在[选项]与[选项]之间，例如[选项]C[选项]，必须给出一个选项。
(**Task description:** Now, based on the following input question, select the correct option from the 4 options. Please fill in the answer between [options] and [options], for example [option] C [option], one option must be given.)

问题 (Question): [Please Put Your Questions Here.]
选项 (Options): [Please Put Your Options Here.]

答案 (Answer):

Figure 11: The Prompt Template for Zero-shot CoT on FinanceIQ Dataset.

Table 14: The prompt template of one-shot CoT on MMLU-Math dataset.

Prompt template of One-shot CoT
Input: # Role: You are an expert in the field of Math. Complete the task provided by the user. # Demonstration: Question: [Please Put Your Question of Demonstration Here] Options: [Please Put Your Options of Demonstration Here] Rationale: [Please Put Your Rationale of Demonstration Here] Correct: [Please Put Your Final Choice of Demonstration Here] # Task Description: According to the following question and options, refer to <# Demonstration>, select one correct option. Please think step by step and follow the answer format in <# Demonstration>, i.e., output the brief chain-of-thought beginning with 'Rationale: ' within 500 characters, and the final choice option beginning with 'Correct: ' from 'A' to 'D'. Question: [Please Put Your Question Here] Options: [Please Put Your Options Here]

Prompt Template for Few-shot CoT using random Demo CoT on FinanceIQ Dataset.

 # Role: 你是一名经济领域的专家，请完成指定的任务。
(**# Role:** You are an expert in the field of economics, please complete the tasks specified by the user.)

 # Demonstration:
[Please Put the random Demonstrations Here.]

 # Initialization
作为<Role>，你可以参考<Demonstration>完成用户给定的任务。
(As a<Role>, you can refer to<Demonstration>to complete the tasks given by the user.)

任务描述：现在根据下列输入的问题，从4个选项中选择一个正确的选项，请将答案填在[选项]与[选项]之间，例如[选项]C[选项]，必须给出一个选项。
(**Task description:** Now, based on the following input question, select the correct option from the 4 options. Please fill in the answer between [options] and [options], for example [option] C [option], one option must be given.)

问题 (Question): [Please Put Your Questions Here.]
选项 (Options): [Please Put Your Options Here.]

答案 (Answer):

Figure 12: The Prompt Template for Few-shot CoT on FinanceIQ Dataset.

Table 15: The prompt template of Re-TASK prompt on MMLU-Math dataset.

Prompt template of Re-TASK
Input: # Role: You are an expert in the field of Math. Complete the task provided by the user. # Knowledge: [Please Put Your Knowledge Here] # Demonstration: Question: [Please Put Your Question of Demonstration Here] Options: [Please Put Your Options of Demonstration Here] Rationale: [Please Put Your Rationale of Demonstration Here] Correct: [Please Put Your Final Choice of Demonstration Here] # Task Description: According to the following question and options, refer to <# Knowledge> and <# Demonstration>, select one correct option. Please think step by step and follow the answer format in <# Demonstration>, i.e., output the brief chain-of-thought beginning with 'Rationale: ' within 500 characters, and the final choice option beginning with 'Correct: ' from 'A' to 'D'. Question: [Please Put Your Question Here] Options: [Please Put Your Options Here]

Table 16: The prompt template of knowledge generation.

Prompt template of Knowledge Generation
Role: You are an expert in the field of Math. Complete the task provided by the user. # Demonstration: ## Question: The hypotenuse of a right triangle measures 10 inches and one angle is 45°. What is the number of square inches in the area of the triangle? ## Knowledge: The area of a right triangle is given by A = (1/2) * base * height. # Task Description: Given the question, please just generate the formula or other knowledge related to the question as brief as possible, like the <# Demonstration>. Just output the one related formula or other knowledge, DO NOT output any other characters. ## Question: [Please Put Your Question Here] ## Knowledge:

Table 17: The prompt template of capability item generation.

Prompt template of Capability Item Generation

Role:
You are an expert in the field of Math. Complete the task provided by the user.

# Demonstration:
{
"question":
"At a certain factory, 10 percent of the staplers produced on Monday were defective and 2 percent of the non-defective staplers were rejected by mistake. If 72 of the non-defective staplers were rejected, what was the number of staplers produced that day?"
"options": [
"A. 4,000",
"B. 4,200",
"C. 4,500",
"D. 4,800"
]
"rationale":
"Step 1. We're told that 10% of staplers in a factory are defective. \n Step 2. X = Total staplers, 0.1X = defective staplers, 0.9X = normal staplers. \n Step 3. We're told that 2% of the normal staplers were rejected by mistake and that this = 72 staplers. \n Step 4. 0.9X(0.02) = 72, 0.018X = 72, 18X = 72,000, X = 4,000.",
"correct": "A"
}

Task Description:
I will give you a piece of knowledge text, please help me generate a four-choice question which is one deduction application of this knowledge, including the question, options, rationale and correct answer.
The knowledge is [Please Put Your Knowledge Here].
The answer is required to follow the **json** format in <# Demonstration>, as:

{
"question": Content of the question,
"options": list of four options,
"rationale": Content of the chain-of-thought with each step starting with 'Step x. ', which is limited to 400 characters,
"correct": the single choice character of correct answer
}
You must and can only generate one deduction example of the given knowledge in the above json format. No extra characters are allowed.

Prompt Template for Re-Task (Lite) on FinanceIQ Dataset.

 # Role: 你是一名金融领域的专家，请通过对<Knowledge Application>的学习加深对响应的<Knowledge Recall>的应用能力，完成用户指定的任务。
(**# Role:** You are an expert in the field of economics, please deepen your application ability of<Knowledge Recall>through learning<Knowledge Application> and complete the tasks specified by the user.)

 # Capability Items for Overall Task:
Knowledge Recall
[Please Put the Knowledge Recall of the overall Task Here.]

Knowledge Application
[Please Put the Knowledge Application of the overall Task Here.]

 # Initialization
作为<Role>，你可以参考<Demonstration>完成用户给定的任务。
(As a<Role>, you can refer to<Demonstration>to complete the tasks given by the user.)

任务描述：现在根据下列输入的问题，从4个选项中选择一个正确的选项，请将答案填在[选项]与[选项]之间，例如[选项]C[选项]，必须给出一个选项。
(**Task description:** Now, based on the following input question, select the correct option from the 4 options. Please fill in the answer between [options] and [options], for example [option] C [option], one option must be given.)

问题 (Question): [Please Put Your Questions Here.]
选项 (Options): [Please Put Your Options Here.]

答案 (Answer):

Figure 13: The Prompt Template for Re-TASK (Lite) on FinanceIQ Dataset.

Prompt Template for Re-Task (Full) on FinanceIQ Dataset.

 # Role: 你是一名金融领域的专家，请通过<Knowledge Example 1>的学习加深对响应的<Knowledge Recall 1>的理解能力，对通过对<Knowledge Application 2>的学习加深对响应的<Knowledge Recall 2>的应用能力，完成用户指定的任务。
(**# Role:** You are an expert in the field of economics, please deepen your understanding of <Knowledge Recall 1> through learning <Knowledge Example 1> and deepen your application ability of <Knowledge Recall> through learning <Knowledge Application> and complete the tasks specified by the user.)

 # Capability Items for Subtasks:
Knowledge Recall 1
[Please Put the Knowledge Recall 1 of Subtasks Here.]

Knowledge Example 1
[Please Put the Knowledge Example 1 of Subtasks Here.]

Knowledge Recall 2
[Please Put the Knowledge Recall 2 of the Subtasks Here.]

Knowledge Application 2
[Please Put the Knowledge Application 2 of the Subtasks Here.]

 # Capability Items for Overall Task:
Knowledge Recall
[Please Put the Knowledge Recall of the overall Task Here.]

Knowledge Application
[Please Put the Knowledge Application of the overall Task Here.]

 # Initialization
作为<Role>，你可以参考<Demonstration>完成用户给定的任务。
(As a<Role>, you can refer to<Demonstration>to complete the tasks given by the user.)

任务描述：现在根据下列输入的问题，从4个选项中选择一个正确的选项，请将答案填在[选项]与[选项]之间，例如[选项]C[选项]，必须给出一个选项。
(**Task description:** Now, based on the following input question, select the correct option from the 4 options. Please fill in the answer between [options] and [options], for example [option] C [option], one option must be given.)

问题 (Question): [Please Put Your Questions Here.]
选项 (Options): [Please Put Your Options Here.]

答案 (Answer):

Figure 14: The Prompt Template for Re-TASK (Full) on FinanceIQ Dataset.

Prompt Template For Generating Knowledge on FinanceIQ Dataset.

 # Role: 你是一名金融领域的专家，请完成用户指定的任务。
(**# Role:** You are an expert in the field of economics, please complete the tasks specified by the user.)

 # Demonstration:
输入(Input): [Please Put a Subtask Here]
输出 (Output): [Please Put the Knowledge of the Subtask Here]

 # Initialization
作为<Role>，你可以参考<Demonstration>完成用户给定的任务。
(As a<Role>, you can refer to<Demonstration>to complete the tasks given by the user.)

任务描述：我将给你一些步骤和知识点，请帮我判断给定的知识点是否有用，如果有用，请帮我总结一个小标题，并给出知识点的具体内容。要求以以下格式输出：1. 知识点1：知识点1的具体内容； 2. 知识点2：知识点2的具体内容。要求只关注我给你的知识点，不要引入任何其他的知识点。只需要按上述格式生成内容，不要输出其他任何字符。
(**Task description:** I will give you some steps and knowledge points. Please help me determine if the given knowledge points are useful. If they are, please summarize a subheading and provide specific content for the knowledge points. Require output in the following format: 1 Knowledge Point 1: The specific content of Knowledge Point 1; 2. Knowledge Point 2: The specific content of Knowledge Point 2. Please only focus on the knowledge points I have given you and do not introduce any other knowledge points. Just generate the content inthe above format and do not output any other characters.)

输入 (Input): [Please Put Your Subtask Here.]

输出 (Output):

Figure 15: The Prompt Template for generating knowledge on FinanceIQ Dataset.

Prompt Template for Generating Capability Items on FinanceIQ Dataset.

 # Role: 你是一名金融领域的专家，请完成用户指定的任务。
(**# Role:** You are an expert in the field of economics, please complete the tasks specified by the user.)

 # Demonstration:
输入(Input): [Please Put the Conceptual Knowledge Here]
输出 (Output): [Please Put An Instances of the input knowledge Here]

输入(Input): [Please Put the Procedural Knowledge Here]
输出 (Output): [Please Put A Application Case of the input knowledge Here]

 # Initialization
作为<Role>，你可以参考<Demonstration>完成用户给定的任务。
(As a<Role>, you can refer to<Demonstration>to complete the tasks given by the user.)

任务描述：我将给你一些知识，一类是程序性知识，一类是概念性知识。请你首先帮我对知识进行分类，对于概念性知识，请你帮我结合实际举一个具体的例子对这个知识进行解释说明，不超过200字。对于程序性知识，请你先帮我结合实际帮我生成一个应用这个程序性知识的选择题的题目、选项和答案，要求答案是先描述解题思路，再生成对应的选项。按照
题目：
选项：
答案：
的格式输出。每个知识点如果是程序性知识，就生成应用题，如果是概念性知识，就生成例子，每个知识点必须生成一个示例，不允许多生成或者少生成。不允许生成多余的字符。
(**Task description:** I will give you some knowledge, one is procedural knowledge, and the other is conceptual knowledge. Please first help me classify the knowledge. For conceptual knowledge, please provide me with a specific example based on practical experience toexplain it, no more than 200 words. For procedural knowledge, please help me generate a multiple-choice question, options, and answers that apply this procedural knowledge based on practical experience. The answer should first describe the problem-solving approach and then generate the corresponding options. according to
Question:
Options:
Answer:
Format output. If each knowledge point is procedural knowledge, generate application questions; if it is conceptual knowledge, generate examples. Each knowledge point must generate one example, and it is not allowed to generate too much or too little. Generating extra characters is not allowed.)

输入 (Input): [Please Put Your Knowledge Here.]

输出 (Output):

Figure 16: The Prompt Template for generating knowledge on FinanceIQ Dataset.