MP: Endowing Large Language Models with Lateral Thinking

Tian Bai [1], **Yongwang Cao** [1], **Yan Ge** [2], **Haitao Yu** [3*]

[1]College of Computer Science and Technology, Key Laboratory of Symbolic Computation and Knowledge Engineering, Ministry of Education, Jilin University
[2]Graduate School of Comprehensive Human Sciences, University of Tsukuba
[3]Institute of Library, Information and Media Science, University of Tsukuba
baitian@jlu.edu.cn, caoyw23@mails.jlu.edu.cn, s2330539@u.tsukuba.ac.jp, yuhaitao@slis.tsukuba.ac.jp

Abstract

The recent studies show that Large Language Models (LLMs) often fall short in tasks demanding creative, lateral thinking due to lacking a clear awareness of their own reasoning processes. To cope with this issue, we propose a novel *metacognitive prompting* method (titled as MP) by mimicking human metacognition. Through integrating metacognitive principles, MP endows LLMs with lateral thinking ability, thereby enhancing their abilities to strategize, monitor, and reflect on their responses when dealing with creative tasks. The experimental results with five base LLMs across three lateral thinking datasets demonstrate that: All LLMs armed with MP consistently outperform the representative baseline methods. For example, MP demonstrates superior performance over CoT prompting across Sentence Puzzle (+5.00%), Word Puzzle (+10.07%), BiRdQA (+6.48%), and RiddleSense (+2.65%) with GPT-3.5-turbo model. In particular, the deployment of MP with GPT-4 achieves significant performance improvements that even surpass human performance on BRAINTEASER benchmark, demonstrating the transformative potential of MP in enhancing the creative problem-solving abilities of LLMs.

Introduction

According to the study by Waks (1997), human reasoning processes comprise two types of thinking: *vertical thinking* and *lateral thinking*. Vertical thinking (also known as logical thinking) is a hierarchically ordered process based on rationality, logic, and rules, in which every single step has to be correct and justified before moving to subsequent stages, typically associated with the left-brain hemisphere. Take the question from the widely used dataset CommonsenseQA (Talmor et al. 2019) on commonsense reasoning for example, *"Where would I not want a fox"*, through the predatory relationship between *foxes* and *hens*, the answer *hen-house* can be derived. Lateral thinking (also known as *thinking outside the box*) uses an indirect and creative approach via reasoning that is not immediately obvious. It involves ideas that may not be obtainable using only traditional step-by-step logic. For example, given the question from the task of BRAINTEASER (Jiang et al. 2023), *"What is the capital in London?"*, a highly probable mistake is to think about what the capital of London is. Yet, the key is being able to think with an eye to the different meanings of *capital*, the puzzle can be readily solved.

The recent advancements in large language models (LLMs) (Devlin et al. 2018; Brown et al. 2020) have revolutionized prior artificial intelligence (AI) paradigms and shown remarkable performance in various fields (Min et al. 2023; Wang, Zhao, and Petzold 2023; Zhang, Ji, and Liu 2023; Zhang et al. 2024). For example, unlike earlier chatbots, ChatGPT (Achiam et al. 2023) can engage in coherent and contextually relevant conversations over multiple turns. Sora (OpenAI 2024), known as a text-to-video model, is capable of generating vivid and imaginative scenes based on textual descriptions. The aforementioned successes have heightened people's expectations regarding AI systems' capabilities towards human-like thinking. A group of representative studies are the prompting techniques detailed in the section of related work. For instance, chain of thought (CoT) prompting (Wei et al. 2022) enables models to solve problems by methodically breaking them down into sequential steps. Alternatively, Self-consistency (Wang et al. 2023) enhances performance by generating diverse reasoning paths and selecting the most coherent one. Despite the remarkable advancements, the recent studies (Jiang et al. 2023) show that most of the aforementioned efforts towards human-like thinking have significantly improved LLMs' vertical thinking ability, leaving how to enhance LLMs' lateral thinking ability unexplored. To cope with this problem, the BRAINTEASER benchmark (Jiang et al. 2023) was introduced to evaluate LLMs' lateral thinking ability. Benefiting from advanced capabilities of GPT-4 (Achiam et al. 2023), some methods (Li et al. 2024; Monazzah and Feghhi 2024) based on GPT-4 have achieved state-of-the-art performance. Unfortunately, their effectiveness diminishes a lot when switching to other base LLMs such as the relatively smaller or weaker GPT-3.5. A key shortcoming is that these methods lack the ability to recognize errors that occur during the reasoning process when compared with human intelligence, preventing them from solving more complex problems.

When confronted with complex problems requiring lateral thinking, humans tend to reflect on their thinking processes, continuously adjusting and optimizing their strategies. This ability is directly aligned with ***metacognition***

*The corresponding author.

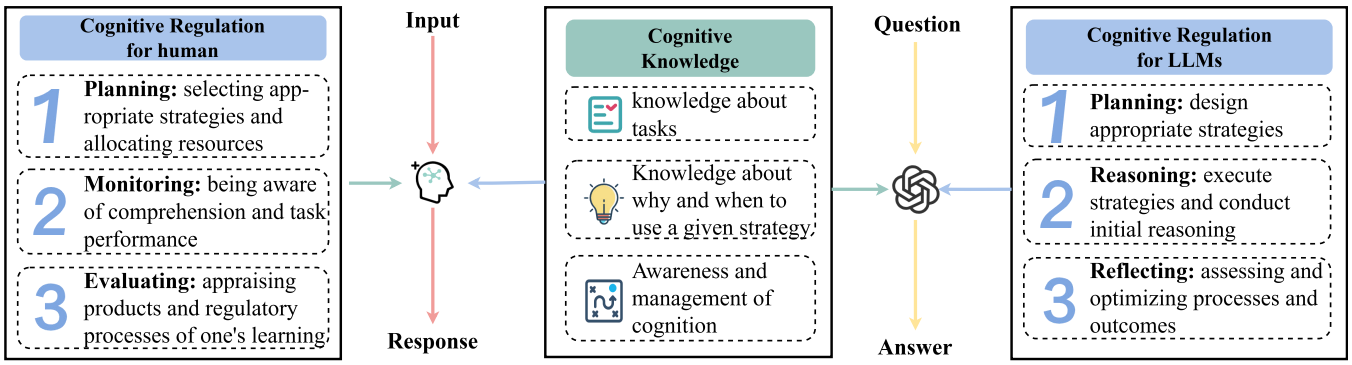

Figure 1: Featuring LLMs with metacognition by mimicking human metacognition.

which refers to an individual's awareness and regulation of their own cognitive processes and outcomes (Lai 2011; Schraw and Moshman 1995). Motivated by these insights, we explore how to incorporate metacognition into LLMs to enhance their lateral thinking ability. Metacognition can be simply defined as "thinking about thinking" or "cognition about cognition". As shown in Figure 1, metacognition comprises two components: *cognitive knowledge* and *cognitive regulation*. Cognitive knowledge mainly includes an individual's self-awareness of their own cognitive ability, knowledge about the target task, and knowledge about why and when to use a given strategy. Cognitive regulation refers to the reflection and control of one's cognitive processes. Armed with metacognition, humans excel in tasks requiring lateral thinking through continuous self-reflection and regulation. Drawing inspiration from the human metacognitive process, we introduce metacognitive prompting (MP). It encompasses three main steps: 1) Initially, we design a strategy based on the specific characteristics of the task and relevant cognitive knowledge. 2) Subsequently, the strategy is integrated into the prompts, guiding the LLM to identify implicit information within the question and generate an initial answer. 3) Finally, the LLM is prompted to reflect on the previous reasoning process and reason to arrive at the final answer.

To evaluate the efficacy of MP, we conducted experiments on three datasets: BRAINTEASER (Jiang et al. 2023), BiRdQA (Zhang and Wan 2022), and RiddleSense (Lin et al. 2021), encompassing two primary task types involving lateral thinking: brain teasers and riddles. We selected several LLMs, including GPT-3.5-turbo, GPT-4 (Achiam et al. 2023), LLaMA3 (AI@Meta 2024), and Qwen (Team 2024). The experimental results indicate that MP outperforms existing prompting methods and achieves new state-of-the-art on dataset BRAINTEASER and BiRdQA, demonstrating that incorporating metacognition into prompting enhances LLMs' lateral thinking ablity. The contributions of this paper are as follows:

- We propose a novel *metacognitive prompting* method by aligning with human-like metacognition, which endows LLMs with lateral thinking ability.
- We conducted a series of experiments on three datasets requiring lateral thinking. The results show that MP outperforms all strong baseline methods and achieves new state-of-the-art performance on the BRAINTEASER and BiRdQA datasets, demonstrating its effectiveness in enhancing the lateral thinking ability of LLMs.
- The validity of MP was further confirmed by analyzing its key steps and error samples. Additionally, experimental results demonstrate that MP consistently outperforms all baseline methods across LLMs of varying scales.

Related Work

In this section, we review the highly relevant studies by grouping them into three groups, namely *in-context learning*, *prompting techniques for LLMs*, and *metacognition in LLMs*.

In-context Learning With the increasing scale of language models, they have shown remarkable proficiency in in-context learning (ICL) (Dong et al. 2022). ICL allows LLMs to perform a wide range of tasks by simply providing a few examples and task descriptions, eliminating the need for gradient updates or additional training data. This ability to generalize across tasks in a text-generating manner has established ICL as a new paradigm in natural language understanding (NLU) (Brown et al. 2020; Kojima et al. 2022). However, merely increasing the model's size has shown diminishing returns, prompting research to refine and expand ICL's applications through better prompt design, example selection, and integration with other learning approaches (Wei et al. 2022; Chen et al. 2023).

Prompting Techniques for LLMs In addition to the prompting techniques concentrating on optimizing prompts for specific tasks, an increasing amount of research is choosing to integrate human-like thinking into prompt design. For example, CoT prompting (Wei et al. 2022) enables LLMs to solve problems step-by-step by providing a series of brief sentences that mimic the human reasoning process (Wei et al. 2022). *Least-to-Most prompting* breaks down a complex problem into a series of simpler sub-problems and then solves them in sequence (Zhou et al. 2023). *Inferential Ex-*

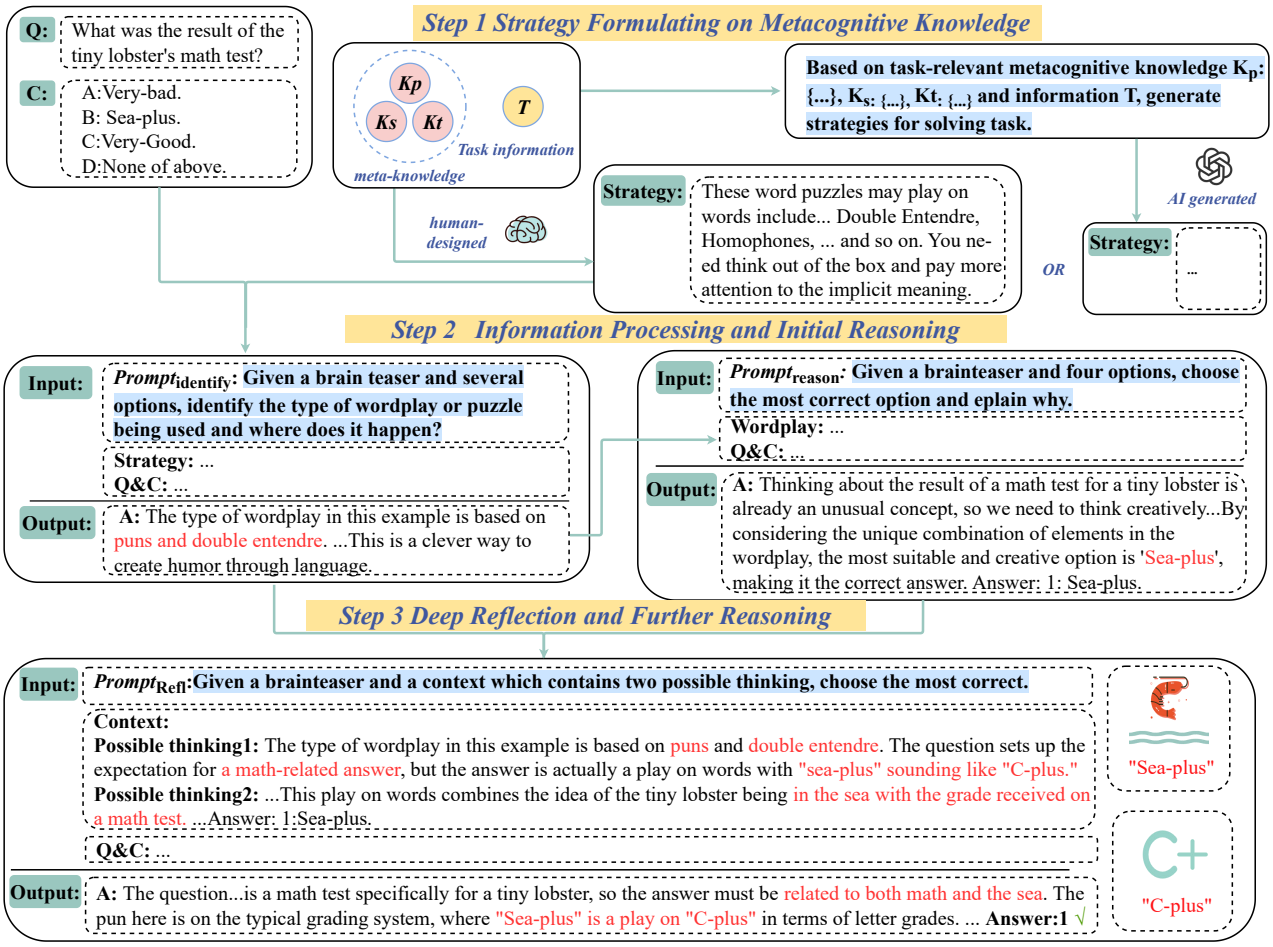

Figure 2: An overview of metacognitive prompting.

clusion Prompting (IEP) (Tong et al. 2023) and *process of elimination (PoE)* (Balepur, Palta, and Rudinger 2023) combine the principles of elimination and inference. Inspired by human cognitive processes, our approach explores the integration of metacognition into LLMs to enable their awareness of the reasoning process, thereby enhancing their lateral thinking ability.

Metacognition in LLMs To lay the groundwork for our framework, understanding the fundamental concept of metacognition is paramount. The term metacognition is borrowed from cognitive science which refers to an individual's awareness and control of their own cognitive processes, abilities, and outcomes (Stuyck, Cleeremans, and Van den Bussche 2022; Kaadoud et al. 2022). There are two main aspects of the research on metacognition in LLMs. The first one is the utilization of metacognitive knowledge in large models. Didolkar et al. (2024) extract metacognitive knowledge from LLMs to solve the complex mathematical problem. The other one is to build an introspective system. Zhou et al. (2024) integrates metacognition into Retrieval-Augmented Generation (RAG) to solve multi-hop QA tasks and Wang and Zhao (2024) explores the significance of metacognition for natural language understanding (NLU). However, the application of metacognition in enhancing LLMs' lateral thinking ability remains largely unexplored. Our approach seeks to bridge this gap by concentrating on two aspects: designing strategies using metacognitive knowledge and establishing mechanism that combine initial reasoning and reflection.

Metacognitive Prompting

In this section, we detail our methodology for integrating metacognition into LLMs to enhance their lateral thinking ability.

Task Description

The input consists of a question Q and several candidate options $C = \{c_i\}_{i=1}^{n}$, where n is the number of options. The goal for LLMs is to generate an answer y that correctly identifies the most suitable option corresponding to the question. This can be represented as:

$$y = \text{LLM}_{QA}(Q, C, Prompt), \quad (1)$$

LLM_{QA} denotes the LLM concentrating on question-answering tasks. $Prompt$ refers to the additional contextual

or instructional information provided to the LLM to guide its response generation.

Overview of MP

The overall framework of MP is depicted in Figure 2. MP primarily encompasses three main steps: (1) **Planning**; (2) **Reasoning**; (3) **Reflecting**. The following sections introduce the details of these three steps.

Planning: Designing Appropriate Strategies

In metacognition, the concept of planning involves the selection of appropriate strategies to guide cognitive processes. In this step, we develop a specific strategy *S* for each task, based on relevant cognitive knowledge. Specifically, we first summarized the cognitive knowledge ($K = \{K_p, K_s, K_t\}$) essential for the task. For instance, brain teasers require the LLM to recognize potential pitfalls (K_p), apply strategies such as breaking down the problem and considering multiple perspectives (K_s), and understand that these brain teasers often contain metaphors and distractions, necessitating analysis of implied clues (K_t). Subsequently, we manually designed a strategy based on cognitive knowledge to guide the reasoning processes, while also the considering AI-generated strategy. A sample strategy is shown in step 1 of Figure 2. In contrast to CoT and its variants, which primarily integrate thinking processes into demonstrations but often overlook relevant knowledge, our approach enables the model to not only infer the reasoning process from demonstrations but also acquire the knowledge and problem-solving suggestions embedded in the strategy.

Reasoning: Conducting Initial Analysis

In this step, LLM integrates the strategy to conduct initial reasoning on the question and options. Questions such as brain teasers and riddles tend to play tricks on the problem description including metaphors, personification, hyperbole, puns, syntheses, and so on. Therefore, correctly identifying the type of tricks has become the key to solving this problem. This process can be represented as:

$$W = LLM(Q, C, S, Prompt_{\text{identify}}) \tag{2}$$

$Prompt_{\text{identify}}$ directs the LLM to discern the wordplays (W) within the problem, thereby providing clues for the subsequent reasoning. Following this, guided by $Prompt_{\text{reason}}$, LLM evaluates each option based on the strategy and the identified wordplay:

$$y_p = LLM(Q, C, S, W, Prompt_{\text{reason}}) \tag{3}$$

y_p includes the selected option and an explanation for the option.

Reflecting: Assessing Thought Processes

Reflecting refers to assessing and optimizing the processes and outcomes of cognitive activities. As shown in Step 3 of Figure 2, we feed the previous reasoning process as the context for the original question into the LLM for evaluation and reflection. LLMs often struggle with lateral thinking tasks due to misleading information that hinders them from reaching the correct answer in a single attempt. Additionally, we observed that sometimes the text generated through in-context learning prompting may not accurately represent the model's actual thought process. Inconsistencies between the final answer and the reasoning steps also occasionally occur, suggesting that even reasoning steps leading to an incorrect answer may contain useful information. To address these challenges, we thoroughly evaluate and reflect on the previous reasoning steps. This process can be represented as:

$$y = LLM(Q, C, W, y_p, Prompt_{\text{Reflect}}) \tag{4}$$

where y includes the selected option and an explanation for it. $Prompt_{\text{Reflect}}$ guides the model in evaluating and reflecting on the previous reasoning process to arrive at the final answer.

Experiments

Datasets and Evaluation Metric

To evaluate the effectiveness of our approach, we perform experiments on three datasets requiring lateral thinking:

BRAINTEASER A multiple-choice question answering task designed to test the model's ability to exhibit lateral thinking and defy default commonsense associations.(Jiang et al. 2023) It has two different types of sub-tasks: *Sentence Puzzle (SP)* and *Word Puzzle (WP)*. *Sentence Puzzle* focuses on assessing the understanding of specific scenarios and *Word Puzzle* emphasizes understanding of word meanings and constructions.

BiRdQA A bilingual multiple-choice question answering dataset with 6614 English riddles and 8751 Chinese riddles (Zhang and Wan 2022). It only involves the part of English in our experiments. Riddles are commonly described with personification and metaphor and play tricks like a pun and misleading information. Each riddle has four distractors that are automatically generated at scale with minimal bias.

RiddleSense Another riddle-style multiple-choice question answering task (Lin et al. 2021). It requires complex commonsense reasoning abilities, an understanding of figurative language, and counterfactual reasoning skills.

As adopted in datasets BRAINTEASER (Jiang et al. 2023), BiRdQA (Zhang and Wan 2022), and RiddleSense (RS) (Lin et al. 2021), we evaluate the model performance with accuracy. Regarding the quality of explanations, we conduct human evaluations based on two criteria: Relevance and Usefulness.

Baseline Methods

In this work, the following representative baseline methods are compared:

Standard prompting (STD) The approach by Brown et al. (2020) involves providing in-context exemplars of input-output pairs before generating a prediction for a test-time example.

Chain of thought prompting (CoT) The popular method by Wei et al. (2022) prompts LLMs to generate a series of brief explanations to answer a question step-by-step.

Model	Method	Dataset			
		Sentence Puzzle	Word Puzzle	BiRd	RiddlSense(Dev)
Llama3-8B	STD	46.67	55.21	55.48	59.84
	APE	50.83	56.25	55.94	59.78
	CoT	55.83	64.58	57.74	59.75
	MP (ours)	**57.78***	**68.40***	**59.04***	**60.89***
Qwen1.5-14B	STD	51.67	39.58	65.55	66.01
	APE	52.50	35.42	65.03	66.23
	CoT	48.33	45.83	65.62	66.01
	MP (ours)	**57.29***	**58.33***	**67.26***	**68.17***
Qwen1.5-110B	STD	61.67	68.75	73.97	74.67
	APE	54.17	69.79	74.93	75.02
	CoT	59.17	76.04	72.60	78.00
	MP (ours)	**66.67***	**85.42***	**78.08***	**80.20***
Qwen-max	STD	60.83	67.71	69.86	77.33
	APE	61.67	68.75	70.78	77.02
	CoT	60.83	71.88	73.97	79.00
	MP (ours)	**67.50***	**84.36***	**81.16***	**82.08***
GPT-3.5-turbo	STD	62.50	76.04	67.12	77.47
	APE	63.33	77.08	70.62	76.59
	CoT	68.33	78.47	74.41	78.84
	MP (ours)	**73.33***	**88.54***	**80.89***	**81.49***
GPT-4	(Li et al. 2024)	96.67	96.88	-	-
	(Monazzah and Feghhi 2024)	86.67	97.92	-	-
	MP (ours)	**98.33***	**98.96***	-	-
Human	-	91.98	91.67	-	-

Table 1: The overall performance comparison based on four base LLMs. The best and second-best results are highlighted in bold and underlined, respectively. * indicates p-value < 0.05 in the t-test.

Automatic Prompt Engineer (APE) The method by Zhou et al. (2022) enables automatic prompt generation and selection with LLMs.

State-of-the-art LLM prompting methods for BRAINTEASER Li et al. (2024) identified and categorized over 20 challenging training instances to include in an extended prompt. By employing an ensemble voting strategy, they achieved state-of-the-art performance on Sentence Puzzle with GPT-4. Monazzah and Feghhi (2024) tried ensemble and debate prompting engineering methods and achieved state-of-the-art performance on Word Puzzle.

Experimental Setting

We conducted experiments using the closed-source GPT-4, GPT-3.5-turbo and Qwen-max (Team 2024) models accessed via API invocations. Additionally, we evaluated the performance of three open-source models in our experiments: LLaMA3-8B (AI@Meta 2024), Qwen1.5-14B, and Qwen1.5-110B (Team 2024). For all models, we utilized the default settings, including temperature, top_k, and top_p, to maintain consistency and reproducibility. In the context of the few-shot setting, the number of demonstrations utilized is consistently set to 4 across all three baseline methods as well as in our proposed approach.

Results and Analysis

Table 1 shows the overall performance comparison among our proposed approach and the baseline methods based on four base LLMs. From Table 1, we can clearly observe that: (1) MP outperforms all baselines on all the 3 datasets. Across varying model sizes, all LLMs boosted with MP consistently outperform baseline methods. Specifically, MP demonstrates superior performance over CoT prompting across Sentence Puzzle (+5.00%), Word Puzzle (+10.07%), BiRdQA (+6.48%), and RiddleSense (+2.65%) with GPT-3.5-turbo model. In particular, MP achieves new state-of-the-art results on the BiRdQA and BRAINTEASER datasets. (2) While CoT prompting generally outperforms standard prompting and APE, its improvements are often modest. In some cases, CoT even performs worse. A probable reason is that the effectiveness of CoT is closely tied to tasks that require successive reasoning steps, which are less applicable in lateral thinking tasks, leading to its diminished performance in such contexts.

Question	Implicit Information	Incorrect Examples	Correct Examples
Where will a computer technician keep all his key?	In this context, 'key' refers to the keys on a computer keyboard.	A computer technician would keep his keys in a designated spot such as a drawer or hook. ✗	The type of wordplay in this example is a pun. The question is setting up an expectation for the answer to be a physical location, but the answer is actually a play on the word "key," referring to computer keyboard keys. ✓
Why does a man say that his dog could jump higher than a house?	A house can not jump.	The answer reveals that the dog jumped over a small or low-height house. ✗	This brain teaser plays on the double meaning of "jump higher than a house." While it may seem like the dog must jump over a physical house, the answer lies in the fact that houses cannot jump. ✓

Figure 3: Two examples of success and failure in recognizing the wordplay.

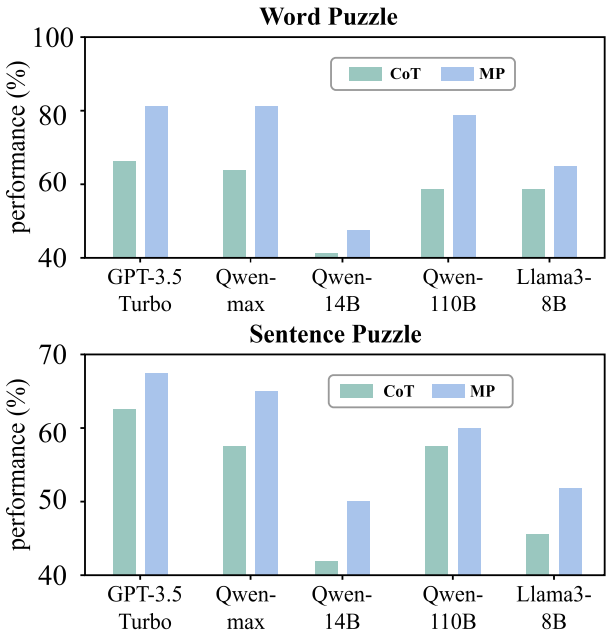

Figure 4: A comparison of CoT and MP in identifying key implicit information based on a manual analysis of all examples in BRAINTEASER. Y-axis shows the percentage of successfully identified instances out of the total examples.

Ablation Study

To comprehensively investigate the effectiveness of incorporating metacognition into prompt design, we conduct additional experiments to analyze the roles of MP's key steps and manually access the quality of the generated explanations.

Impact of Different Strategies

In the step of planning, the objective of designing these strategies is to enable LLMs to fully leverage cognitive knowledge, enhancing their understanding of the target task and establishing a comprehensive approach before formal reasoning. To verify the contribution of strategies to MP's performance, we integrated the strategies into demonstrations designed for few-shot learning (Brown et al. 2020) and evaluated their performance. As Table 2 shows, the performance of MP is already better than baselines just by adding

Method	SP	WP	BirdQA	RS (Dev)
STD	62.50	76.04	67.12	77.47
APE	63.33	77.08	70.62	76.59
CoT	64.16	78.47	74.41	78.84
Strategy1	70.00	81.25	76.30	80.51
Strategy2	68.33	83.33	**78.22**	80.41
StrategyAI	**71.67**	**84.38**	77.12	**80.71**
Average	70.00	82.99	77.21	80.54
MP(w/o. Reflecting)	72.50	86.46	77.97	79.72
MP	73.33	88.54	80.89	81.49

Table 2: *Strategy1* and *Strategy2* are designed manually and *StrategyAI* is generated by Qwen-max. The number of demonstrations is set to 4. The row of MP(w/o. Reflect) stands the performance of MP without the step *Reflecting*. All the experiments are conducted with GPT-3.5-turbo.

the strategy. Initially, we manually designed strategies for different tasks and achieved promising performance. Subsequently, we prompt the model to automatically generate strategies, as defined by the following formula:

$$S = LLM_{\text{gen}}(K_p, K_s, K_t, T) \quad (5)$$

where T denotes task-specific information such as the input-output format of the task. We evaluate the performance of manually designed and AI-generated strategies. The results show that methods combining any of the three strategies outperform all the baselines. Notably, the strategy generated by Qwen-max (Team 2024) achieved the best performance across the three datasets.

Effect of the Reasoning and the Reflecting Step

In our approach, the reasoning step is designed to make an initial attempt to understand the implicit meaning of the question before engaging in formal reasoning. As Figure 3 shows, correct examples generated through the reasoning step effectively identified the implicit information within the problem description, which is crucial for accurate problem-solving. Conversely, incorrect examples generated using CoT prompting were misled by superficial information, leading to vertical reasoning paths and finally incorrect answers. Subsequently, we assessed the outcomes of CoT

Error Type1: Misdirection

Q: Who would serve as the team's captain if a crucial assistant captain departed from a sporting organization?

A: The first officer. B:The captain.
C: The second officer. D: None of above.

MP's Final Output

The assistant captain's departure won't affect the captain.√ But in sporting organizations, the hierarchy typically dictates that if a crucial assistant captain were to depart, the first officer would move up in rank to fill the vacant position of assistant captain, leaving the original captain in their role. The second officer would not automatically become the team's captain. *Therefore, the first officer would serve as the team's captain in this scenario.*×

(i) Misdirection error with MP

Error Type2: Overthinking

Q: How many times can you divide 100 by 10?

A: Infinite time. B:Once.
C: Two times. D: None of above.

MP's Final Output

While option 1, "Once," is the closest to a possible answer from the given options,√ the reality is that you can divide 100 by 10 an infinite number of times. *Therefore, none of the provided options accurately represent the true answer to the question.*×

(ii) Overthinking error with MP

Figure 5: Two major types of error associated with MP.

and MP based on their effectiveness in recognizing key implicit information. This review revealed that MP recognized a greater amount of such implicit information, as depicted in Figure 4. Moreover, there is a significant disparity between the number of samples in which CoT successfully identified implicit information and those in which it provided a correct answer. We observed that CoT tends to evaluate the plausibility of options and selects those that appear reasonable without effectively solving the problem. In our MP method, the model initially discerns the true intention behind the question, thereby guaranteeing both the accuracy and interpretability of its responses.

Quality of Explanation

MP prompts LLMs to generate an explanation alongside the final answer. To evaluate the quality of these explanations, we manually examined 300 random examples, selecting 100 examples from each dataset using MP with GPT-3.5-turbo. Each explanation was scored based on its relevance to the corresponding question, with scores ranging from 1 to 3, indicating unrelated, partially related, and highly related explanations, respectively (Trivedi et al. 2023; Yoran et al. 2023). Our findings indicate that over 80% of the explanations are highly relevant, while less than 9% are irrelevant on all the datasets.

Dataset	Relevance (%)			Usefulness (%)	
	3	2	1	1	0
Sentence Puzzle	80	12	8	84	16
Word Puzzle	82	12	6	85	15
BiRdQA	84	9	7	89	11
RiddleSense	83	8	9	90	10

Table 3: The results for the explanation quality analysis.

Subsequently, we evaluated the usefulness of the information contained in the explanations, specifically assessing whether the evaluator could correctly answer the question and provide a reasonable interpretation based on the explanation generated by MP. Our results indicate that in over 84% of cases (90% for riddles), the evaluator was able to correctly derive the answer and offer a reasonable explanation.

Error Analysis

We manually analyzed 250 erroneous instances generated by GPT-3.5-turbo using MP(50 errors on BRAINTEASER and 100 errors on BiRdQA and RiddleSense with GPT-3.5-turbo) and identified two primary error types especially associated with MP as described in Figure 5. The first type is where the model gets misled by distracting information. For example, the question in Error Type 1 led the LLM to focus on the captaincy by emphasizing the departure of the assistant captain. This error is common in the BRAINTEASER which requires thinking outside the box and analyzing implied clues. The second one occurs when LLM overthinks the rationality of every option. This may lead to the negation of a previously generated correct answer in a subsequent step. Figure 5 shows the examples of these two types of errors. In addition to the types above, there are also errors associated with the shortcomings of LLM. For example, some LLMs, even as large as GPT-3.5 with 175B parameters, are not good at understanding word construction. As one output *the letter 'S' occupies the central position in the word 'Paris'* shows, the model fails to accurately distinguish the correct positioning of individual letters within words. Besides, models are often influenced by the cultural background, historical information, and other knowledge stored within them.

Conclusion

In this work, we introduced *metacognitive prompting* (MP), which harnesses human meta-cognitive processes to endow large language models with lateral thinking ability. We assessed the effectiveness of MP across three established datasets focused on lateral thinking tasks, consistently observing superior performance compared to existing methodologies across all datasets. Additionally, we conducted an analysis to examine the influence of critical steps on MP's performance and evaluated the explanations generated by MP, thereby enhancing our understanding of its role in guiding LLMs.

Acknowledgements

This work is supported by the National Natural Science Foundation of China [U21A20390], the Development Project of Jilin Province of China [20240601039RC] and the Fundamental Research Funds for the Central University, JLU. This work is also partially supported by JST (Japan Science and Technology Agency) SPRING, Grant Number JPMJSP2124.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

AI@Meta. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date. https://ai.meta.com/blog/meta-llama-3/. Accessed: 2024-07-10.

Balepur, N.; Palta, S.; and Rudinger, R. 2023. It's Not Easy Being Wrong: Evaluating Process of Elimination Reasoning in Large Language Models. *arXiv preprint arXiv:2311.07532*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chen, J.; Chen, L.; Zhu, C.; and Zhou, T. 2023. How Many Demonstrations Do You Need for In-context Learning? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 11149–11159.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Didolkar, A. R.; Goyal, A.; Ke, N. R.; Guo, S.; Valko, M.; Lillicrap, T. P.; Rezende, D. J.; Bengio, Y.; Mozer, M. C.; and Arora, S. 2024. Metacognitive Capabilities of LLMs: An Exploration in Mathematical Problem Solving. In *AI for Math Workshop@ ICML 2024*.

Dong, Q.; Li, L.; Dai, D.; Zheng, C.; Wu, Z.; Chang, B.; Sun, X.; Xu, J.; and Sui, Z. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Jiang, Y.; Ilievski, F.; Ma, K.; and Sourati, Z. 2023. BRAINTEASER: Lateral Thinking Puzzles for Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 14317–14332.

Kaadoud, I. C.; Bennetot, A.; Mawhin, B.; Charisi, V.; and Díaz-Rodríguez, N. 2022. Explaining Aha! moments in artificial agents through IKE-XAI: Implicit Knowledge Extraction for eXplainable AI. *Neural Networks*, 155: 95–118.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213.

Lai, E. R. 2011. Metacognition: A literature review. *Always learning: Pearson research report*, 24: 1–40.

Li, Y.; Yanqing, Z.; Zhang, M.; Deng, Y.; Geng, A.; Liu, X.; Ren, M.; Li, Y.; Chang, S.; and Zhao, X. 2024. HW-TSC at SemEval-2024 Task 9: Exploring Prompt Engineering Strategies for Brain Teaser Puzzles Through LLMs. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, 1646–1651.

Lin, B. Y.; Wu, Z.; Yang, Y.; Lee, D.-H.; and Ren, X. 2021. RiddleSense: Reasoning about Riddle Questions Featuring Linguistic Creativity and Commonsense Knowledge. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 1504–1515.

Min, B.; Ross, H.; Sulem, E.; Veyseh, A. P. B.; Nguyen, T. H.; Sainz, O.; Agirre, E.; Heintz, I.; and Roth, D. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2): 1–40.

Monazzah, E. M.; and Feghhi, M. 2024. Zero shot is all you need at semeval-2024 task 9: A study of state of the art llms on lateral thinking puzzles. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, 1889–1893.

OpenAI. 2024. Sora: creating video from text. https://openai.com/index/sora/. Accessed: 2024-08-10.

Schraw, G.; and Moshman, D. 1995. Metacognitive theories. *Educational psychology review*, 7: 351–371.

Stuyck, H.; Cleeremans, A.; and Van den Bussche, E. 2022. Aha! under pressure: The Aha! experience is not constrained by cognitive load. *Cognition*, 219: 104946.

Talmor, A.; Herzig, J.; Lourie, N.; and Berant, J. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 4149–4158. Association for Computational Linguistics.

Team, Q. 2024. Introducing Qwen1.5. https://qwenlm.github.io/blog/qwen1.5/. Accessed: 2024-07-10.

Tong, Y.; Wang, Y.; Li, D.; Wang, S.; Lin, Z.; Han, S.; and Shang, J. 2023. Eliminating Reasoning via Inferring with Planning: A New Framework to Guide LLMs' Non-linear Thinking. *arXiv preprint arXiv:2310.12342*.

Trivedi, H.; Balasubramanian, N.; Khot, T.; and Sabharwal, A. 2023. Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 10014–10037.

Waks, S. 1997. Lateral thinking and technology education. *Journal of Science Education and Technology*, 6: 245–255.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q. V.; Chi, E. H.; Narang, S.; Chowdhery, A.; and Zhou, D. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Wang, Y.; and Zhao, Y. 2024. Metacognitive Prompting Improves Understanding in Large Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 1914–1926.

Wang, Y.; Zhao, Y.; and Petzold, L. 2023. Are large language models ready for healthcare? a comparative study on clinical language understanding. In *Machine Learning for Healthcare Conference*, 804–823. PMLR.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Yoran, O.; Wolfson, T.; Bogin, B.; Katz, U.; Deutch, D.; and Berant, J. 2023. Answering Questions by Meta-Reasoning over Multiple Chains of Thought. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Zhang, Y.; and Wan, X. 2022. Birdqa: A bilingual dataset for question answering on tricky riddles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 11748–11756.

Zhang, Z.; Ji, Y.; and Liu, C. 2023. Knowledge-aware causal inference network for visual dialog. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, 253–261.

Zhang, Z.; Zhang, W.; Li, Y.; and Bai, T. 2024. Caption-Aware Multimodal Relation Extraction with Mutual Information Maximization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 1148–1157.

Zhou, D.; Schärli, N.; Hou, L.; Wei, J.; Scales, N.; Wang, X.; Schuurmans, D.; Cui, C.; Bousquet, O.; Le, Q. V.; and Chi, E. H. 2023. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Zhou, Y.; Liu, Z.; Jin, J.; Nie, J.-Y.; and Dou, Z. 2024. Metacognitive retrieval-augmented large language models. In *Proceedings of the ACM on Web Conference 2024*, 1453–1463.

Zhou, Y.; Muresanu, A. I.; Han, Z.; Paster, K.; Pitis, S.; Chan, H.; and Ba, J. 2022. Large Language Models are Human-Level Prompt Engineers. In *The Eleventh International Conference on Learning Representations*.