# Metacognitive Prompting Improves Understanding in Large Language Models

Yuqing Wang<sup>1</sup>, Yun Zhao<sup>2</sup>

<sup>1</sup>Stanford University, <sup>2</sup>Meta Platforms

NAACL 2024

2024.05.21

Presenter: Hawon Jeong

## Introduction

- The nuanced understanding abilities of LLM remain unexplored
- Metacognition: "Thinking about thinking"
  - 메타인지: 자기 인지 능력, 자신이 무엇을 알고 무엇을 모르는지 아는 것, 자신의 사고과정에 대한 이해 및 평가가 가능
- Metacognitive Prompting:

This approach integrates key aspects of human metacognitive process into LLM

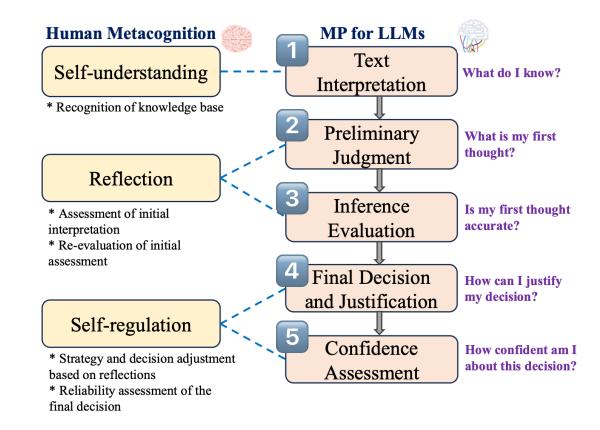


Figure 1: Alignment between human metacognitive processes and the stages of MP in LLMs.

## **Related Works**

### Prompting Techniques in LLMs

- Current research focuses on "reasoning abilities"
- E.g., CoT, Self-Consistency, Least-to-Most, ToT, and Plan and Solve Prompting
- But enhancing NLU remains challenge

## Natural Language Understanding in NLP

- NLU, the fundamental aspect of NLP, is a model's capacity to grasp semantics and nuances of human language
- The inherent NLU competencies of LLMs have remained relatively inadequately explored

## Cognitive Processes in NLU

- Cognitive processes heavily influence our linguistic abilities
- In domain of NLU, incorporating cognitive insights may offer improvements

— Overview

## Human's high-level cognition is from...

- 1. Breaking down abstract concepts
- 2. Critically evaluating scenarios
- 3. Finetuning our reasoning.

→ Let's equip LLMs that simulates
the self-reflective cognitive process!

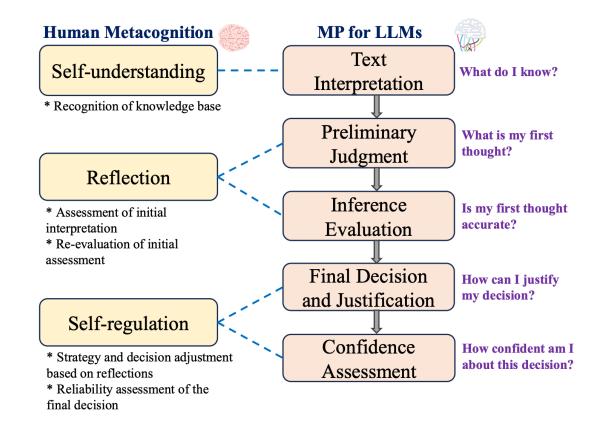
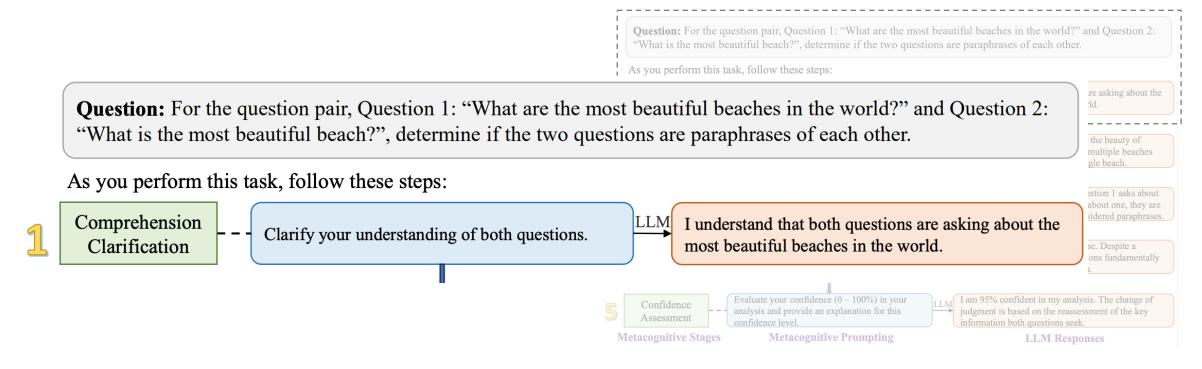


Figure 1: Alignment between human metacognitive processes and the stages of MP in LLMs.

### Overview

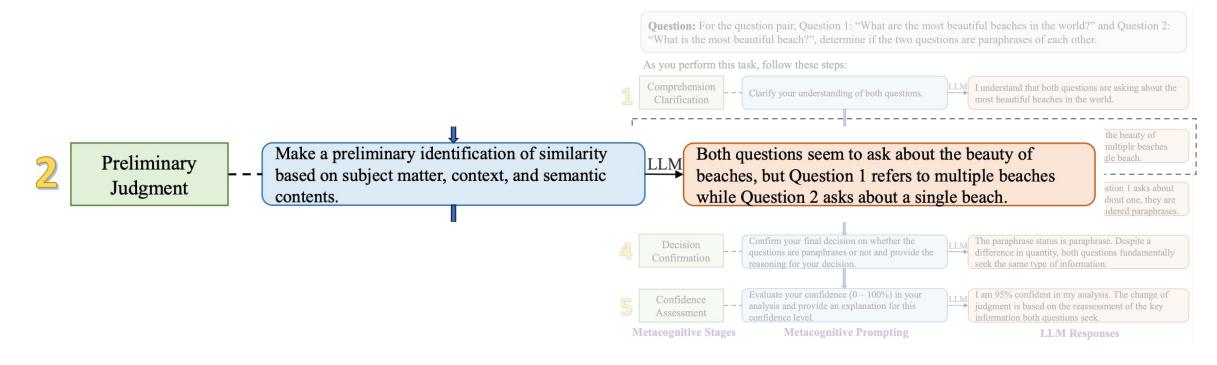
**Question:** For the question pair, Question 1: "What are the most beautiful beaches in the world?" and Question 2: "What is the most beautiful beach?", determine if the two questions are paraphrases of each other. As you perform this task, follow these steps: Comprehension I understand that both questions are asking about the Clarify your understanding of both questions. most beautiful beaches in the world. Clarification Make a preliminary identification of similarity Both questions seem to ask about the beauty of **Preliminary** LLM based on subject matter, context, and semantic beaches, but Question 1 refers to multiple beaches Judgment while Question 2 asks about a single beach. contents. Critically assess your preliminary analysis. If you Upon reassessment, although Question 1 asks about 3 Critical are unsure about the initial assessment of the multiple beaches and Question 2 about one, they are **Evaluation** questions being paraphrases, try to reassess it. essentially similar and hence considered paraphrases. Confirm your final decision on whether the The paraphrase status is paraphrase. Despite a Decision difference in quantity, both questions fundamentally questions are paraphrases or not and provide the Confirmation reasoning for your decision. seek the same type of information. Evaluate your confidence (0 - 100%) in your I am 95% confident in my analysis. The change of Confidence LLM analysis and provide an explanation for this judgment is based on the reassessment of the key Assessment confidence level. information both questions seek. **Metacognitive Stages Metacognitive Prompting LLM Responses** 

— Step 1. Comprehension Clarification



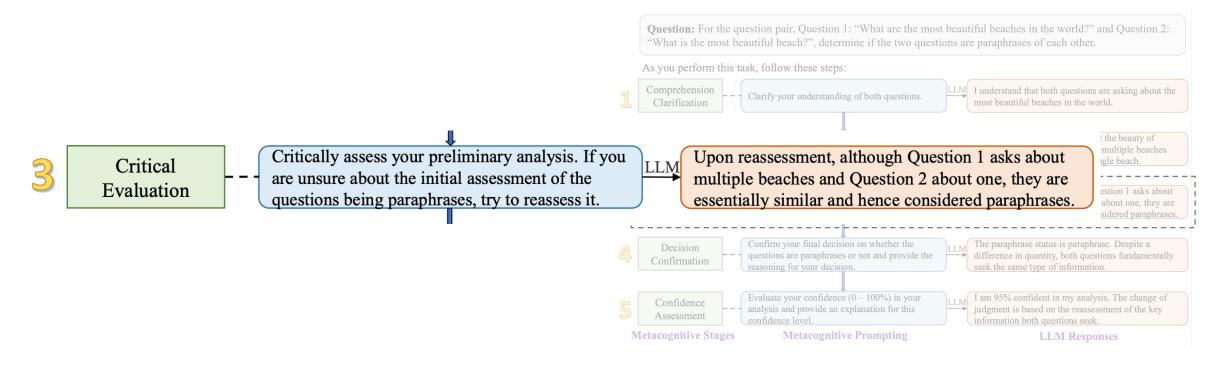
- LLM begins by deciphering the input text to comprehend its context and meaning Logical Thinking
- This is mirroring the initial comprehension stage in human thought

— Step 2. Preliminary Judgement



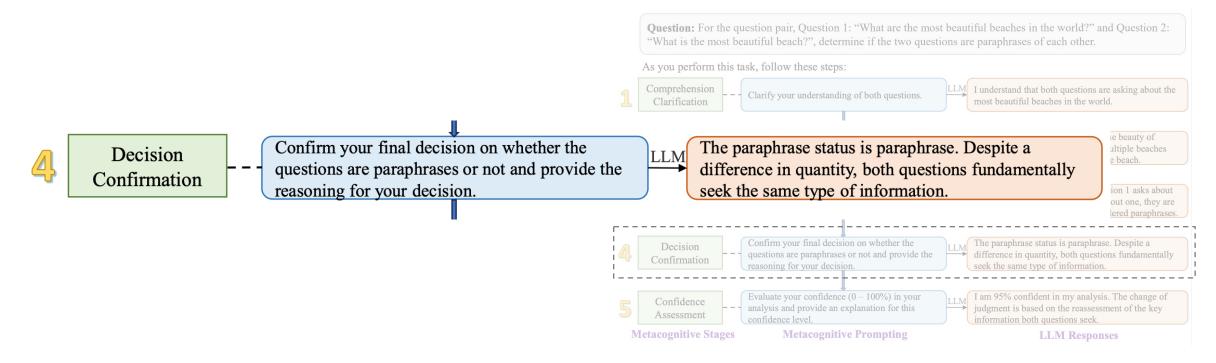
- LLM then forms a preliminary interpretation of the text
- This is a step that reflects judgment formation in humans

— Step 3. Critical Evaluation



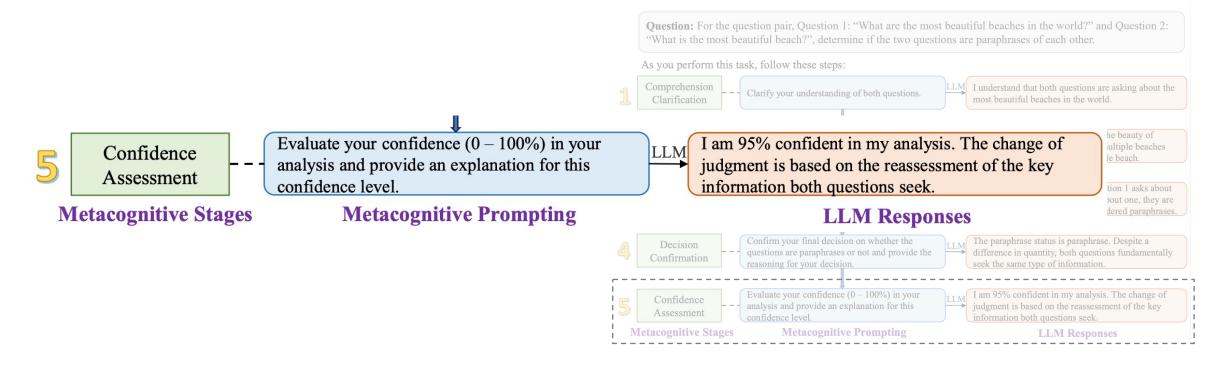
- LLM critically evaluates this initial judgment for accuracy
- This is akin to the self-scrutiny humans apply during problem-solving

— Step 4. Decision Confirmation



- LLM finalizes its decision and offers an explanation for its reasoning
- This aligns with the decision-making and rationalization phase in human cognition

Step 5. Confidence Assessment



- LLM assesses its confidence in the outcome of the entire process
- This is similar with how humans gauge the certainty of their decisions and explanation

# **Experiments**

### — Datasets and Models

Source Benchmark	Dataset	Task	# Classes	Metrics	Domain	
GLUE	QQP QNLI	Paraphrase QA/NLI	<ul><li>2 (paraphrase or not)</li><li>2 (entailment or not)</li></ul>	acc./F1 acc.	Social QA Wikipedia	
SuperGLUE	BoolQ WiC	QA WSD	2 (yes/no) 2 (True/False)	acc.	Wikipedia, Google queries WordNet, Wiktionary, etc.	
BLUE	BC5CDR-chem DDI MedNLI	NER RE NLI	3 (BIO tags) 4 (Advice, Effect, etc.) 3 (ECN relations)	μ-F1 m-F1 acc.	Biochemistry Biochemistry Clinical practice	
LexGLUE	EUR-LEX LEDGAR UNFAIR-ToS	MLC MCC MLC	100 (EuroVoc concepts) 100 (contract provisions) 8 + 1 (unfair terms)	μ-F1/m-F1 μ-F1/m-F1 μ-F1/m-F1	EU Law Contracts Contracts	

### • LLMs

• LLaMA-2-13B-Chat, PaLM-2-Bison-Chat, GPT-3.5-Turbo, GPT-4

# **Experiments**

— Datasets, Prompts, and Models

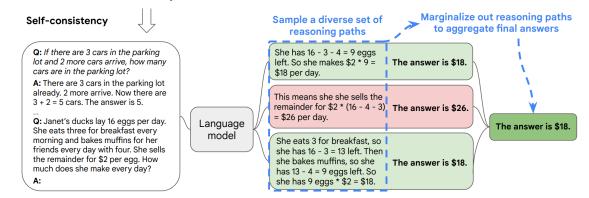
### Prompts

### 1) Zero-shot baselines

- Zero-shot CoT: "Lets think step by step."
- Plan-and-Solve Prompting: "Let's first understand the problem and devise a plan to solve the problem. Then, let's carry out the plan and solve the problem step by step."

### 2) Few-shot baselines

- Manual-CoT
- Self-consistency with CoT



#### **Model Input**

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

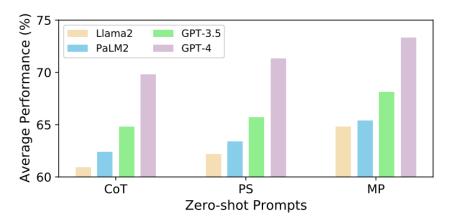
A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

## — Overall Performance Comparison

	Dataset									
Method	QQP	QNLI	BoolQ	WiC	BC5CDR-chem	DDI	MedNLI	EUR-LEX	LEDGAR	UNFAIR-ToS
	acc./F1	acc.	acc.	acc.	$\mu ext{-}FI$	m- $F1$	acc.	$\mu$ -F1/m-F1	$\mu$ -F1/m-F1	$\mu$ -F1/m-F1
Llama2 (0S, CoT)	84.5/79.5	89.5	81.9	75.2	94.2	70.5	58.3	25.6/14.5	60.8/47.6	43.9/26.7
Llama2 (0S, PS)	85.6/80.8	89.9	83.1	76.0	95.6	72.0	59.1	27.8/16.9	61.4/48.1	46.1/28.4
Llama2 (0S, MP)	86.9/82.1	90.4	86.3	78.8	96.0	74.3	62.8	32.5/21.4	63.8/50.5	50.2/31.6
PaLM2 (0S, CoT)	85.4/80.6	89.9	88.1	76.4	94.5	70.9	61.1	24.8/13.1	63.9/49.1	46.2/29.1
PaLM2 (0S, PS)	85.2/80.3	89.5	89.5	77.1	94.9	72.8	60.9	26.1/14.8	65.0/52.7	47.4/30.8
PaLM2 (0S, MP)	86.2/81.9	90.8	90.5	78.8	96.2	74.0	63.3	29.3/16.5	67.6/54.8	52.5/33.7
GPT-3.5 (0S, CoT)	84.9/79.9	90.3	84.8	76.9	93.9	63.9	70.6	31.9/20.7	68.1/57.6	50.4/33.2
GPT-3.5 (0S, PS)	84.7/80.6	90.8	85.0	76.6	94.2	66.1	72.3	33.6/21.8	68.9/58.3	52.3/34.8
GPT-3.5 (0S, MP)	86.1/81.5	92.3	87.7	78.4	94.8	70.7	76.4	36.7/23.5	70.2/59.8	56.7/38.1
GPT-4 (0S, CoT)	88.9/84.7	95.0	90.4	82.0	97.3	72.1	78.2	37.4/24.8	73.6/59.4	54.7/38.5
GPT-4 (0S, PS)	89.4/85.3	96.2	90.7	82.4	97.6	73.5	79.8	39.6/27.1	75.4/60.7	58.3/41.7
GPT-4 (0S, MP)	89.9/86.2	<u>97.1</u>	<u>91.4</u>	<u>83.6</u>	<u>98.5</u>	<u>74.7</u>	<u>81.1</u>	43.8/29.9	78.1/62.8	64.0/45.3
Llama2 (5S, M-CoT)	85.2/80.2	90.1	82.8	76.5	94.9	73.8	61.2	23.3/12.7	54.7/43.3	52.8/35.6
Llama2 (5S, CoT-SC)	86.1/80.9	90.8	84.2	76.9	95.3	76.2	63.5	24.6/14.7	55.6/44.8	55.6/37.9
Llama2 (5S, M-MP)	88.1/83.2	91.6	87.4	79.5	96.6	77.3	64.7	27.8/15.9	58.2/46.6	59.7/41.2
PaLM2 (5S, M-CoT)	85.8/81.3	90.9	89.2	77.7	95.1	73.1	63.3	22.8/12.0	57.5/45.2	57.4/31.9
PaLM2 (5S, CoT-SC)	86.9/81.7	91.7	90.9	78.2	96.4	75.4	63.8	23.9/13.8	57.9/45.7	60.2/34.6
PaLM2 (5S, M-MP)	87.9/82.5	93.8	90.9	79.6	96.2	75.2	65.1	26.7/15.4	59.3/47.3	65.4/38.8
GPT-3.5 (5S, M-CoT)	85.1/80.2	91.2	86.7	77.4	94.7	67.8	74.3	29.3/19.5	61.7/50.1	62.3/45.1
GPT-3.5 (5S, CoT-SC)	86.1/81.7	91.4	88.3	78.8	95.7	70.1	76.5	30.6/19.8	63.0/51.4	65.7/47.2
GPT-3.5 (5S, M-MP)	86.4/81.9	93.1	89.7	79.1	96.6	71.6	78.1	32.4/20.7	64.9/53.7	69.1/50.1
GPT-4 (5S, M-CoT)	89.5/85.6	95.8	90.8	82.3	97.9	74.6	80.1	35.3/22.6	66.4/57.2	69.2/50.3
GPT-4 (5S, CoT-SC)	90.1/86.7	96.8	91.6	83.4	98.9	76.9	80.5	37.6/24.4	68.2/58.4	72.8/54.1
GPT-4 (5S, M-MP)	91.3/88.2	98.9	92.0	84.3	99.4	80.8	82.4	40.1/28.8	70.3/59.9	75.6/55.8

## — Prompting Strategy Comparison



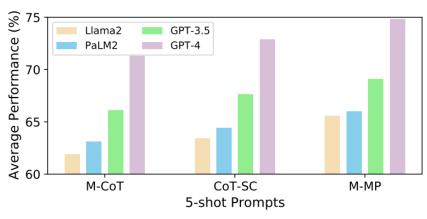


Table 3: Comparison of average performance for zeroshot prompting methods across datasets. Performance metrics are averaged over all models. MP consistently achieves superior performance across all NLU tasks.

Dataset	CoT	PS	MP
QQP (acc./F1)	85.9/81.2	86.2/81.7	87.3/82.9
QNLI (acc.)	91.2	91.6	92.6
BoolQ (acc.)	86.3	87.1	89.0
WiC (acc.)	77.6	78.0	<b>79.9</b>
BC5CDR-chem ( $\mu$ -F1)	95.0	95.6	96.4
DDI(m-F1)	69.4	71.1	73.4
MedNLI (acc.)	67.1	68.0	70.9
EUR-LEX ( $\mu$ - $F1/m$ - $F1$ )	29.9/18.3	31.8/20.2	35.6/22.8
LEDGAR ( $\mu$ -F1/m-F1)	66.6/53.4	67.7/54.9	69.9/57.0
UNFAIR-ToS (μ-F1/m-F1)	48.8/31.9	51.0/33.9	55.8/37.2

Model level comparison

Data level comparison

## — Error Analysis

• Two main types of errors with MP across 10 datasets with manually inspection of incorrect prediction.

## **Error Type 1: Overthinking Question:**

68.3%

In two sentences, sentence 1: "Excite the neurons." sentence 2: "The fireworks which opened the festivities excited anyone present.", both contain the target word "excite". Determine if the target word is used with the same meaning in both sentences. [MP Instructions]

#### MP Initial Response (Stage 2):

The target word has the same meaning: False.



#### MP Final Response (Stage 4 & 5):

The target word has the same meaning: True. Upon reevaluation, model hypothesized that both contexts hint at triggering a reaction, either biological or emotional, leading to the conclusion of similarity. Confidence in this revised analysis is 90%.

(a) Overthinking error in model response with MP.

## **Error Type 2: Overcorrection Question:**

31.7%

In two sentences, sentence 1: "The very easiness of the deed held her back." sentence 2: "There was an easiness between them.", both contain the target word "easiness". Determine if the target word is used with the same meaning in both sentences. [MP Instructions]

#### **MP Initial Response (Stage 2):**

The target word has the same meaning: False.



#### MP Final Response (Stage 4 & 5):

The target word has the same meaning: True. Upon reevaluation, 'easiness' in both sentences pertains to emotional states—implying simplicity-induced hesitation in the first and emotional harmony in the second. Confidence in this revised analysis is 85%.

(b) Overcorrection error in model response with MP.

- Domain-specific task
  - Biomedical NLU tasks: 'Terminological misalignments' (48.6%), 'Clinical inference discrepancies' (51.4%)
  - Legal NLU tasks: 'Statutory interpretation errors' (52.2%), 'Jurisprudential analysis deviations' (47.8%)

— Confidence Analysis

### Observation

- High TP, Low TN: Reliable self-awareness and selfassessment
- FP, FN: Pointing potential improvements

### High confidence, if > 75%

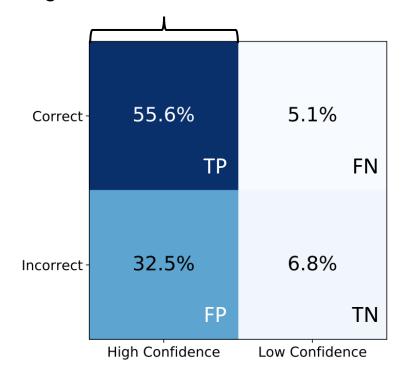


Figure 5: The relationship between correctness and confidence levels under MP, averaged over all datasets and models.

## **Limitations and Discussion**

### Limitation

- 1. Designing metacognitive prompts requires manual effort
- 2. Selection of datasets and models
- 3. The verbalized confidence might not serve as definite confidence method

→ Suggestion: Combining verbalization with self-consistency checks?

### Discussion

- Applying MP more broadly... e.g., arithmetic, mental health care...
- Introducing introspective LLMs, particularly regarding biases and the reliability of outputs