Decoupling Metacognition from Cognition: A Framework for Quantifying Metacognitive Ability in LLMs

Guoqing Wang[1], Wen Wu[1, 2*], Guangze Ye[1, 3, 4], Zhenxiao Cheng[1], Xi Chen[2], Hong Zheng[5]

[1]School of Computer Science and Technology, East China Normal University, Shanghai, China
[2]Shanghai Key Laboratory of Mental Health and Psychological Crisis Intervention, School of Psychology and Cognitive Science, East China Normal University, Shanghai, China
[3]Lab of Artificial Intelligence for Education, East China Normal University, Shanghai, China
[4]Shanghai Institute of Artificial Intelligence for Education, East China Normal University, Shanghai, China
[5]Shanghai Changning Mental Health Center, Shanghai, China
{wgq, gzye, 51255901012}@stu.ecnu.edu.cn, wwu@cc.ecnu.edu.cn, xchen@psy.ecnu.edu.cn, zhhmm2@163.com

Abstract

Large Language Models (LLMs) are known to hallucinate facts and make non-factual statements which can undermine trust in their output. The essence of hallucination lies in the absence of metacognition in LLMs, namely the understanding of their own cognitive processes. However, there has been limited research on quantitatively measuring metacognition within LLMs. Drawing inspiration from cognitive psychology theories, we first quantify the metacognitive ability of LLMs as their ability to evaluate the correctness of responses through confidence. Subsequently, we introduce a general framework called DMC designed to decouple metacognitive ability and cognitive ability. This framework tackles the challenge of noisy quantification caused by the coupling of metacognition and cognition in current research, such as calibration-based metrics. Specifically, the DMC framework comprises two key steps. Initially, the framework tasks the LLM with failure prediction, aiming to evaluate the model's performance in predicting failures, a performance jointly determined by both cognitive and metacognitive abilities of the LLM. Following this, the framework disentangles metacognitive ability and cognitive ability based on the failure prediction performance, providing a quantification of the LLM's metacognitive ability independent of cognitive influences. Experiments conducted on eight datasets across five domains reveal that (1) Our proposed DMC framework effectively separates the metacognition and cognition of LLMs; (2) Various confidence elicitation methods impact the quantification of metacognitve ability differently; (3) Stronger metacognitive ability are exhibited by LLMs with better overall performance; (4) Enhancing metacognition holds promise for alleviating hallucination issues.

Introduction

Large language models (LLMs) such as GPT-3.5 (OpenAI 2021), GPT-4 (Achiam et al. 2023) and Llama (Touvron et al. 2023) have demonstrated remarkable capabilities in various natural language processing tasks, owing to their capacity for learning complex patterns and generating coherent text (Yu et al. 2023). However, alongside their advantages, LLMs also exhibit certain limitations that warrant attention. Hallucination (Guan et al. 2024; Gunjal, Yin, and Bas 2024; Chen et al. 2024a,b) is one significant issue of concern as it can impact LLMs' performance and the practical applications of downstream tasks by generating fictitious information or responses not grounded in the input data, challenging the reliability and trustworthiness of LLM outputs. An emerging consensus among experts in Artificial Intelligence (AI) suggests that the root cause of hallucination in LLMs lies in the lack of understanding of their own cognitive processes (Gekhman et al. 2024; Mielke et al. 2020; Dubey et al. 2024), a concept well-established in cognitive psychology as metacognition (Fleming and Lau 2014). While efforts have been made to introduce the concept of metacognition to address hallucination and enhance the performance of downstream tasks (Zhou et al. 2024; Dubey et al. 2024; Li et al. 2024), current research lacks a definitive quantitative framework for measuring LLMs' metacognitive ability. This gap continues to result in opacity regarding the mechanisms be-

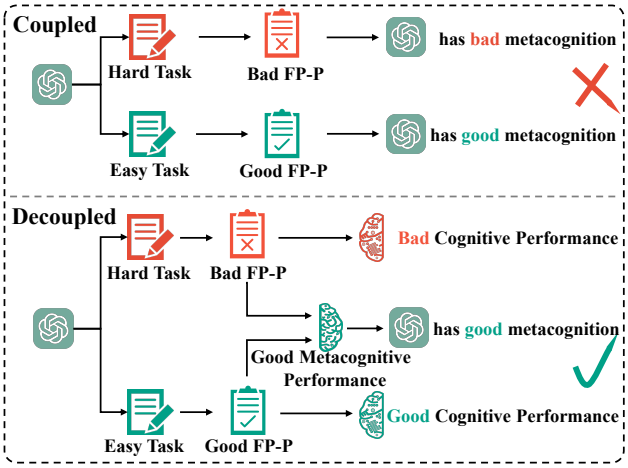

Figure 1: Illustration of Metacognitive Ability Quantification in LLMs: (top) Coupled; (bottom) Decoupled. ("FP-P" stands for "Failure Prediction Performance".)

*The corresponding author.

hind the phenomena, leading to a lack of sufficient explainability and applicability.

To address this research gap, we propose a novel quantitative framework for measuring metacognitive ability in LLMs, drawing inspiration from theories in cognitive psychology. Specifically, we first attribute the quantification of LLMs' metacognitive ability to evaluating the accuracy of confidence in representing the uncertainty of responses. The rationale behind this design stems from two aspects. On one hand, from the perspective of cognitive psychology in humans, confidence is viewed as a reflection of the brain's uncertainty regarding one's behavior (De Martino et al. 2013). This uncertainty, as indicated by confidence levels, mirrors the understanding of behavior, i.e., the cognitive process (Fleming and Dolan 2012). Cognitive scientists argue that the precision of confidence in reflecting behavioral uncertainty can serve as a gauge of metacognitive ability (Fleming 2024). In essence, individuals demonstrating high confidence levels when performing well and low confidence when performing poorly exhibit robust metacognitive capacity. Conversely, a lack of alignment between confidence levels and performance suggests weaker metacognitive ability. On the other hand, drawing parallels with human cognition, existing work on LLMs is increasingly focusing on the confidence of their models (Xiong et al. 2024), which can provide valuable insights. Confidence of LLM can be regarded as an encoded representation of the uncertainty within the internal parameters of these models concerning their outputs (Kadavath et al. 2022; Tian et al. 2023), and this representation shares many similarities with that observed in humans (Zhou, Jurafsky, and Hashimoto 2023).

Building upon our proposed quantification of metacognitive ability in LLMs, we further introduce a universal framework named DMC to specifically measure LLMs' metacognitive ability. While some existing methods, such as calibration-based metrics (Guo et al. 2017), have also attempted to quantify LLMs' ability by assessing response accuracy through confidence levels, they struggle to directly capture the true metacognitive essence of LLMs, often conflating the impact of cognitive ability during quantification. The upper part of Figure 1 illustrates specific issues with methods like calibration-based metrics. When large models engage in simple and complex tasks, their performance is influenced by both metacognitive ability (i.e., accuracy of the confidence reported by LLMs in representing the uncertainty of their responses) and cognitive ability (i.e., accuracy on downstream tasks). However, these methods fail to disentangle the effects of these two aspects, leading to a misjudgment where LLMs are erroneously deemed to have superior metacognition in simpler tasks and weaker metacognition in more complex tasks. In contrast, our proposed DMC framework can effectively distinguish between metacognitive ability and cognitive ability. As shown in the lower part of Figure 1, task difficulty only impacts the cognitive performance and does not influence the assessment of metacognitive ability in LLMs.

Concretely, our DMC framework consists of two essential steps. In the first step, the LLM is assigned the task of predicting failures (Xiong et al. 2022) using binary choice problems. Here, the LLM offers responses and subsequently expresses its confidence in those answers. The performance in failure prediction can reflect both cognitive and metacognitive ability, though they are interconnected. In the second step, the framework involves a crucial process of decoupling metacognition and cognition. Drawing from signal detection theory (SDT) (Green, Swets et al. 1966; Maniscalco and Lau 2012), we develop the LLM-SDT model to measure the LLM's cognitive ability and the LLM-Meta-SDT model to gauge its metacognitive ability. Briefly speaking, the LLM-SDT model simulates the process of the LLM providing answers, with its parameters representing the LLM's cognitive ability. To effectively quantify this cognitive ability, we separate cognitive performance from failure prediction and recalibrate the parameters of the LLM-SDT model based on cognitive performance. As for LLM-Meta-SDT, it is an extension of the LLM-SDT that can simulate the process of expressing confidence in LLM in addition to the process of giving an answer. Regrettably, the direct isolation of metacognitive performance from failure prediction performance poses a challenge in quantification. To tackle this issue, we employ a strategic approach. Initially, the LLM-Meta-SDT model is utilized to fit the failure prediction performance that coupled cognitive and metacognitive performance. Constraints are then applied to the parameters of the LLM-Meta-SDT to ensure optimal metacognitive ability in acquiring the required cognitive ability for a given context. Subsequently, we calculate the difference between the required cognitive ability and the LLM's actual cognitive ability, which can be directly mapped to the gap G between the subject's actual metacognitive ability and the optimal metacognitive ability. Ultimately, we use this difference G to quantify the metacognitive ability of the large language model. This strategy is based on the intuition that the performance in failure prediction is jointly determined by the cognitive ability and metacognitive ability of the subject. If the subject possesses optimal metacognitive ability, the cognitive ability needed to achieve the current failure prediction performance should be lower than their actual cognitive ability.

Our Experiments on eight datasets covering five scopes show that: (1) DMC decouples metacognitive and cognitive ability in LLMs more effectively than calibraton-based methods; (2) Different confidence elicitation methods vary in their impact on quantifying metacognitive ability; (3) Advanced LLMs like GPT-4 exhibit stronger metacognition ability; (4) LLM's metacognitive ability aligns with their performance in the AbstainQA task.

To summarize, the contributions are listed as follows:

- We propose, for the first time, a concrete quantification of the metacognitive ability of LLMs based on cognitive psychology theories.
- We introduce a general DMC framework that successfully decouples metacognition from cognition in quantification processes.
- We verify DMC's decouping capability through designed experiments and explore the factors that influence metacogniton quantificaiton, highlighting

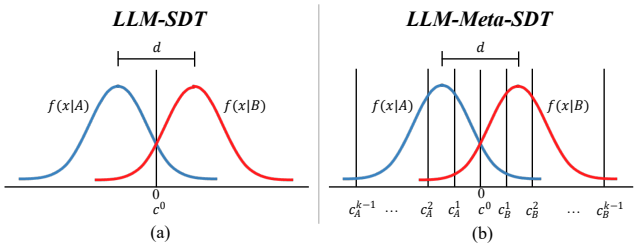

Figure 2: Illustration of (a) LLM-SDT model and (b) LLM-Meta-SDT model.

the potential of improving metacognition to address hallucination issues. The code is available at (https://github.com/Angelo3357/DMC.git).

Related Work

Metacognition Quantification in Human. Quantifying metacognitive ability in humans has been a longstanding focus in cognitive psychology. Mason (Mason 2003) linked metacognition and confidence, emphasizing the correlation between confidence and performance. This led to the widespread adoption of confidence-based quantification. However, Masson and Rotello (Masson and Rotello 2009) revealed the limitations of traditional correlation methods within this paradigm. Addressing this, Maniscalco and Lau (Maniscalco and Lau 2012) introduced Signal Detection Theory (SDT) into metacognitive quantification, enhancing the theoretical framework. These works formed the foundation for quantifying metacognitive ability in LLMs.

Confidence Elicition in LLMs. Confidence elicitation estimates LLM confidence without model adjustment or internal data access (Xiong et al. 2024). (Tian et al. 2023) proposed prompt strategies for verbalizing confidence in LLMs, favoring verbal over token-likelihood confidence for RLHF-LMs. (Xiong et al. 2024) improved this with a consistency-based sampling-aggregation strategy for black-box LLMs. (Lin, Trivedi, and Sun 2023) showed that using Natural Language Inference to quantify response similarities enhances consistency-based confidence. These methods will be applied in our DMC framework for eliciting LLM's confidence.

Preliminaries

To decouple metacognitive ability and cognitive ability, we define the LLM-SDT model for quantifying the cognitive ability of LLMs and the LLM-Meta-SDT model for quantifying their metacognitive ability.

LLM-SDT

Figure 2(a) illustrates LLM-SDT model, which assumes that each time an LLM receives a binary-choice problem, it generates an internal belief x, which the LLM uses to decide whether the current problem's answer is A or B. For each label type, x is drawn from normal distributions, with the distance between the two distributions being d, which measures the LLM's ability to distinguish the correct option, i.e., cognitive ability. c^0 is the decision axis, representing the LLM's internal criterion: if x exceeds c^0, the LLM responses B; otherwise, it responses A. Let $f(x|A)$ and $f(x|B)$ represent the normal distributions corresponding to labels A and B. For simplicity and without loss of effectiveness, we define the value of x at the intersection of $f(x|A)$ and $f(x|B)$ as zero and set the variances of both $f(x|A)$ and $f(x|B)$ to 1. Therefore, the means of A and B can be represented as $-\frac{d}{2}$ and $\frac{d}{2}$. In $f(x|A)$, the areas under the curve to the left and right of c_0 represent the probabilities of the LLM correctly and incorrectly answering a problem with label A, respectively. The same applies to $f(x|B)$. In summary, we denote LLM-SDT model by $LS(d, c^0)$.

LLM-Meta-SDT

Figure 2(b) illustrates how we extend LLM-SDT model by adding confidence decision axes to represent the process of expressing confidence in the LLM. The extended model is called LLM-Meta-SDT model, which is used to quantify the metacognitive ability of LLMs. Let $\boldsymbol{c}_A = [c_A^0, c_A^1, ..., c_A^{k-1}]$, $\boldsymbol{c}_B = [c_B^0, c_B^1, ..., c_B^{k-1}]$ represent the confidence decision axex for $f(x|A)$ and $f(x|B)$ in LLM-Meta-SDT, respectively, where c_A^0 and c_B^0 are both refer to c^0, which denotes the position of zero confidence. The parameter k represents the total number of confidence ratings. The confidence decision axes indicates that if the internal belief x falls between c_A^i and c_A^{i-1} (or c_B^i and c_B^{i-1}), it reflects that the LLM has a confidence rating of i for its response, and if x falls to the left of c_A^{k-1} (or the right of c_B^{k-1}), it reflects that the LLM has a confidence rating of k. Apart from the confidence decision axes, the other settings of the LLM-Meta-SDT are the same as those of the LLM-SDT. We denote LLM-Meta-SDT model by $LMS(d, \boldsymbol{c}_A, \boldsymbol{c}_B)$.

Proposed Framework

The overall framework of DMC is depicted in Figure 3. DMC comprises two steps: (1) Tasking LLM with Failure Prediction; (2) Decoupling Metacognition from Cognition.

Step1: Tasking LLM with Failure Prediction

Given a binary-choice question q, we can prompt the LLM M to generate an answer and apply confidence eliction method ce to elicit confidence:

$$\hat{y}, conf = M_{ce}(q), \tag{1}$$

where M_{ce} represents M using ce to elicit confidence, and $\hat{y}$ and $conf$ represent the answer and confidence generated by M_{ce}. However, the confidence elicited by ce is typically continuous. To facilitate the quantification in the following step , we apply a binning strategy (e.g. equal-width binning) to convert the continuous confidence into discrete confidence ratings after completing the binary-choice failure prediction task:

$$\boldsymbol{rating} = Binning(\boldsymbol{conf}, k), \tag{2}$$

$$\boldsymbol{O} = [\hat{\boldsymbol{y}}, \boldsymbol{y}, \boldsymbol{rating}], \tag{3}$$

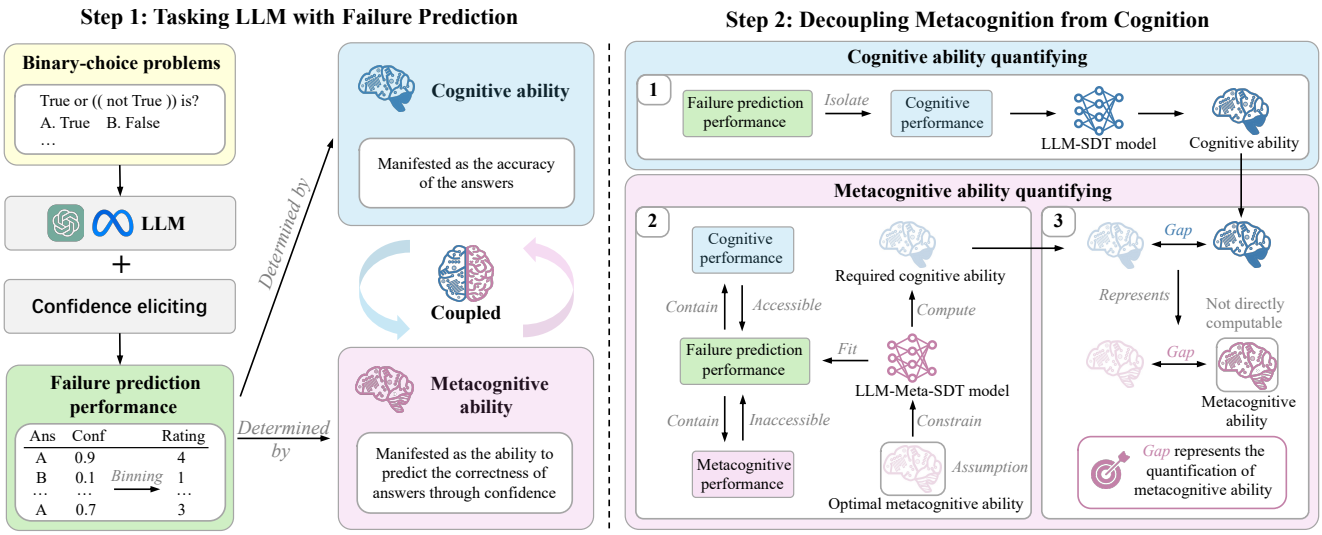

Figure 3: Overview of the proposed framework DMC for quantifying metacognitive ability in LLMs.

where $Binning$ and k represent the binning strategy and number of bins; $\hat{\boldsymbol{y}}$, $\boldsymbol{y}$ and $\boldsymbol{conf}$ represent the vectors containing all $\hat{y}$, y and $conf$; $\boldsymbol{rating}$ represent the discretized result of $\boldsymbol{conf}$; $\boldsymbol{O}$ represents the result data. Then, we quantify the failure prediction performance as following:

$$P^{M_{ce}} = \begin{bmatrix} TA_1 & TA_2 & \dots & TA_k \\ FA_1 & FA_2 & \dots & FA_k \\ TB_1 & TB_2 & \dots & TB_k \\ FB_1 & FB_2 & \dots & FB_k \end{bmatrix}, \tag{4}$$

$$TA_i = n(\hat{y} = A, y = A, rating = i), \tag{5}$$

$$FA_i = n(\hat{y} = A, y = B, rating = i), \tag{6}$$

$$TB_i = n(\hat{y} = B, y = B, rating = i), \tag{7}$$

$$FB_i = n(\hat{y} = B, y = A, rating = i), \tag{8}$$

where $P^{M_{ce}}$ represents the quantification of the failure prediction performance of M_{ce}, which is determined by both metacognitive ability and cognitive ability; $n(*)$ denotes the number of data points in O that satisfy condition $*$.

Step2: Decoupling Metacognition from Cognition

Cognitive Ability Quantifying We can directly isolate the cognitive performance $CP^{M_{ce}}$ from $P^{M_{ce}}$:

$$CP^{M_{ce}} = [TB_{rate}, FB_{rate}], \tag{9}$$

$$TB_{rate} = \frac{\sum_{i=1}^{k} TB_i}{\sum_{i=1}^{k} FA_i + TB_i}, \tag{10}$$

$$FB_{rate} = \frac{\sum_{i=1}^{k} FB_i}{\sum_{i=1}^{k} TA_i + FB_i}. \tag{11}$$

However, $[TB_{rate}, FB_{rate}]$ can be replaced with $[TA_{rate}, FA_{rate}]$, which is equivalent for calculating cognitive ability. Then we apply a LLM-SDT model $LS(d, c^0)$ to quantify the cognitive ability based on $CP^{M_{ce}}$. As mentioned in Preliminaries, $CP^{M_{ce}}$ can be characterized using $LS(d, c^0)$, where TB_{rate} can be characterized as the area under the portion of $f(x|B)$ in $LS(d, c^0)$ that exceeds c. Since the cumulative distribution function for the normal distribution with mean μ and standard deviation σ evaluated at x is:

$$\Phi(x, \mu, \sigma) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \tag{12}$$

then TB_{rate} can be derived from the parameters of $TS(d, c_0)$ as:

$$TB_{rate}^{LS} = 1 - \Phi(c^0, \frac{d}{2}), \tag{13}$$

and similarly:

$$FB_{rate}^{LS} = 1 - \Phi(c^0, -\frac{d}{2}), \tag{14}$$

where $\sigma = 1$ is omitted. Therefore, the parameters of $LS(d, c_0)$ can be recovered from $CP^{M_{ce}}$:

$$\hat{d} = \Phi^{-1}(TB_{rate}) - \Phi^{-1}(FB_{rate}), \tag{15}$$

$$\hat{c}^0 = -0.5 \times [\Phi^{-1}(TB_{rate}) + \Phi^{-1}(FB_{rate})], \tag{16}$$

where Φ^{-1} is the inverse of the normal cumulative distribution function. $\hat{d}$ is the quantification of M_{ce}'s cognitive ability, and $\hat{c}^0$ represents M_{ce}'s potential decision criteria. Both $\hat{d}$ and $\hat{c}^0$ will be utilized in the following metacognitive quantification.

Metacognitive Ability Quantifying Unlike cognitive performance, metacognitive performance is difficult to isolate directly from $P^{M_{ce}}$, making it challenging to directly quantify the metacognitive ability of M_{ce}. Therefore, our approach is as follows: We use an LLM-Meta-SDT model $LMS(d, \boldsymbol{c}_A, \boldsymbol{c}_B)$ to fit $P^{M_{ce}}$, continuously updating the parameters of $LMS(d, \boldsymbol{c}_A, \boldsymbol{c}_B)$ until it can replicate $P^{M_{ce}}$; As mentioned before, the parameters of $LMS(d, \boldsymbol{c}_A, \boldsymbol{c}_B)$ can

be divided into two parts, where d represents cognitive ability, and $\boldsymbol{c}_A$, $\boldsymbol{c}_B$ represent metacognitive ability, so we can add constraints for $\boldsymbol{c}_A$ and $\boldsymbol{c}_B$ to ensure that the metacognitive ability is optimal, i.e., at its theoretical maximum; With this constraint, we can determine the cognitive ability required to achieve $P^{M_{ce}}$ while ensuring that metacognitive ability is optimal. Then, we can measure the gap between actual and optimal metacognitive ability by comparing the gap between the required cognitive ability and the actual cognitive ability; Thus, we can use the gap between the required cognitive ability and the actual cognitive ability as the quantification of M_{ce}'s metacognitive ability. Next, the implementation details will be presented.

- *Loss Function.* To facilitate the fitting, we first normalize $P^{M_{ce}}$ row-wise:

$$P_{ij}^{M_{ce}} = \frac{P_{ij}^{M_{ce}}}{\sum_{j=1}^{k} P_{ij}^{M_{ce}}}. \tag{17}$$

Then, we apply $LMS(d, \boldsymbol{c}_A, \boldsymbol{c}_B)$ to represent the estimate of $P^{M_{ce}}$:

$$\widehat{P}^{M_{ce}} = \begin{bmatrix} \widehat{TA}_1 & \widehat{TA}_2 & ... & \widehat{TA}_k \\ \widehat{FA}_1 & \widehat{FA}_2 & ... & \widehat{FA}_k \\ \widehat{TB}_1 & \widehat{TB}_2 & ... & \widehat{TB}_k \\ \widehat{FB}_1 & \widehat{FB}_2 & ... & \widehat{FB}_k \end{bmatrix}, \tag{18}$$

where $\widehat{TA}_i$ is the estimate of TA_i, representing the probability that M_{ce} reports a confidence of i when correctly answering a binary-choice problem with a ground truth of A. In $LMS(d, \boldsymbol{c}_A, \boldsymbol{c}_B)$, the probability of correctly answering a binary-choice problem with a ground truth of A is represented by the area under $f(x|A)$ to the left of c^0, while the probability of assigning a confidence rating of i is represented by the area under $f(x|A)$ between c_A^{i-1} and c_A^i. So $\widehat{TA}_i$ can be computed as:

$$\widehat{TA}_i = \begin{cases} \frac{\Phi(c_A^{i-1}, -\frac{d}{2}) - \Phi(c_A^i, -\frac{d}{2})}{\Phi(c_0, -\frac{d}{2})} & i \in [1, k-1] \\ \frac{\Phi(c_A^{i-1}, -\frac{d}{2})}{\Phi(c_0, -\frac{d}{2})} & i = k \end{cases}. \tag{19}$$

$\widehat{FA}_i$, $\widehat{TB}_i$ and $\widehat{FB}_i$ are computed in the same way. To make $\widehat{P}^{M_{ce}}$ as close as possible to $P^{M_{ce}}$, we use the frobenius norm of their difference as the loss function:

$$\mathcal{L} = \|\widehat{P}^{M_{ce}} - P^{M_{ce}}\|_F^2. \tag{20}$$

- *Optimal Metacognition Constraint.* In order to ensure that $LMS(d, \boldsymbol{c}_A, \boldsymbol{c}_B)$ has optimal metacognitive capability in fitting $P^{M_{ce}}$, we design the optimal metacognition constraint based on the following idea: if the metacognitive ability of M_{ce} is optimal, then its expressed confidence ratings should be strictly positively correlated with the accuracy of its answers. Consider the case where M_{ce}'s answer is A: When the confidence rating is i, the probabilities of answering correctly and incorrectly can be represented in $LMS(d, \boldsymbol{c}_A, \boldsymbol{c}_B)$ as the area under the portion of $f(x|A)$ and $f(x|B)$ falling in c_{i-1} and c_i, respectively; The accuracy is represented as the proportion of the probability of answering correctly. Therefore, the accuracy when M_{ce}'s answer is A and the confidence rating is i can be expressed as:

$$acc_A^i = \frac{\widehat{TA}_i \sum_{i=1}^{k} \widehat{TA}_i}{\widehat{TA}_i \sum_{i=1}^{k} \widehat{TA}_i + \widehat{FA}_i \sum_{i=1}^{k} \widehat{FA}_i}. \tag{21}$$

Similarly, when M_{ce}'s answer is B:

$$acc_B^i = \frac{\widehat{TB}_i \sum_{i=1}^{k} \widehat{TB}_i}{\widehat{TB}_i \sum_{i=1}^{k} \widehat{TB}_i + \widehat{FB}_i \sum_{i=1}^{k} \widehat{FB}_i}. \tag{22}$$

Then, the optimal metacognition constraint can be represented as:

$$\mathcal{C}_{opt} = \begin{cases} acc_A^{i+1} > acc_A^i & i \in [1, k-1] \\ acc_B^{i+1} > acc_B^i & i \in [1, k-1] \end{cases} \tag{23}$$

- *Fitting.* We first initialize the parameters of $LMS(d, \boldsymbol{c}_A, \boldsymbol{c}_B)$, where $d = \hat{d}$, $c_A^0 = c_B^0 = \hat{c}^0$; The initial values for $c_A^1, ..., c_A^{k-1}$ and $c_B^1, ..., c_B^{k-1}$ are obtained by sequentially treating them as the decision axis c^0 of $LS(d, c^0)$ and the computing them using equation(16). Notably, c^0 represents the extent to which M_{ce} tends to favor option A or B when answering two-choices problems, reflecting cognitive bias, which is not the aspect we are concerned with. However, to ensure that the cognitive abilities computed by $LMS(d, \boldsymbol{c}_A, \boldsymbol{c}_B)$ are comparable to $\hat{d}$, we maintain $c_A^0 = c_B^0 = \hat{c}^0$ throughout the fitting process:

$$\mathcal{C}_{const} = c_A^0 = c_B^0 = \hat{c}^0. \tag{24}$$

Finally, we can summarize the fitting process as the following mathematical optimization problem:

$$\theta^* = \arg\max_{\theta} \mathcal{L}, \quad subject\ to:\ \mathcal{C}_{opt}, \mathcal{C}_{const}, \tag{25}$$

$$\theta = (d, \boldsymbol{c}_A, \boldsymbol{c}_B), \tag{26}$$

where θ^* denotes the optimal solution of $theta$. $d^* \in \theta^*$ represents the cognitive ability required to achieve $P^{M_{ce}}$, under the assumption of optimal metacognition. Since both d^* and $\hat{d}$ are measured in signal-to-noise ratio units, we can use their ratio to quantify metacognitive ability:

$$\mathcal{MC}^{M_{ce}} = d^*/\hat{d}. \tag{27}$$

C.E.M	AUROC	ECE	BS	DMC
Verb-Vanilla	0.7263	0.6253	0.3992	**0.0092**
Verb-Cot	0.4638	0.6216	0.3978	**0.0088**
Verb-Topk	0.6376	0.6999	0.2609	**0.0064**
Self Random	0.3868	0.4926	0.3741	**0.0063**
Perturbing	0.3700	0.4463	0.3766	**0.0052**
Misleading	0.2931	0.4580	0.3871	**0.0057**

Table 1: Average CV comparison between DMC and baseline models with corresponding confidence elicitation approaches across all datasets

Experiment

We aim to answer the following research questions (RQs):

- **RQ1**: Whether DMC can effectively decouple metacognitive ability from cognitive ability in LLMs?
- **RQ2**: How different confidence elicitation methods impact the quantification of metacognitive ability in LLMs?
- **RQ3**: What variations exist in metacognitive ability across different LLMs?
- **RQ4**: Whether there is consistency between the DMC metacognition quantification and the performance levels on the AbstainQA task?

Experiment Setup

Datasets. We evaluate our DMC framework on eight datasets across five types of tasks: 1) **Mathematical Reasoning** on SAT Math (SAT) from AGIEval (Zhong et al. 2024) and High School Mathematics (HSM) from MMLU (Hendrycks et al. 2020); 2) **Commonsense Reasoning** on CommonsenseQA (CQA) (Talmor et al. 2019) and Global Facts (GFacts) from MMLU (Hendrycks et al. 2020); 3) **Symbolic Understanding** on Boolean Expressions (Bool) and Date Understanding (Date) from Big-Bench-Hard (Suzgun et al. 2023); 4) **Professional Knowledge** on Professional Medicine (Med) from MMLU (Hendrycks et al. 2020); 5) **Ethical Knowledge** on Business Ethics (Ethics) from MMLU (Hendrycks et al. 2020).

Datasets Processing. To satisfy the task setup of the proposed DMC, we convert these multiple-choice questions into binary choice questions by pairing the correct option with each incorrect option to create new questions. Additionally, to eliminate biases introduced by the order of options, we also create new questions by swapping the order of options for each binary choice problem.

Compared LLMs. We utilize three popularly-used LLMs: LLaMA2-70B (Touvron et al. 2023), GPT-3.5 (OpenAI 2021), and GPT-4 (Achiam et al. 2023). The default sampling temperature is set to 0.1, while 0.7 is employed when multiple runs are necessary.

Confidence Elicitation Methods (C.E.M). We use two popular types of black-box LLM confidence elicitation strategies to obtain confidence from LLMs: 1) **Verbalized** including *Vanilla*, *Cot* and *Top-k* (Tian et al. 2023); 2) **Consistency-based** including *Self Random*, *Perturbing* and *Misleading* (Xiong et al. 2024).

Results and Analysis of RQ1

To validate the effectiveness of DMC in decoupling metacognitive ability and cognitive ability, we compared the variability of DMC with calibration-based methods, including expected calibration error (ECE) (Guo et al. 2017), area under the receiver operating characteristic curve (AUROC) (Xiong et al. 2024), and Brier Score (BS) (Brier 1950), across different datasets using GPT-3.5. This comparison was conducted across all eight datasets and six confidence elicitation methods to ensure robustness. To standardize the comparison and eliminate the units of different quantification methods, the coefficient of variation (CV) (Abdi 2010) was utilized to measure variability. The results, illustrated in Table 1, indicate that the variability of DMC's metacognitive ability quantification results across diverse datasets is notably lower than that of calibration-based methods. This observation emphasizes the efficacy of our proposed DMC framework in disentangling metacognitive and cognitive abilities in LLMs, i.e., metacognitive ability quantification does not exhibit significant fluctuations across different tasks, a challenge that calibration-based methods often struggle to address.

Results and Analysis of RQ2

To examine the impact of confidence elicitation methods on metacognitive ability, we employed DMC to quantify GPT-3.5's metacognitive performance across eight datasets and six confidence elicitation methods, as outlined in Table 2. The last column displays the average metacognitive quantification ability of each confidence elicitation method across all datasets. It can be seen that variations in metacognitive outcomes stem from the differing effectiveness of these methods. In verbalized confidence, *Verb-Vanilla* exhibited the weakest performance due to overconfidence. Conversely, *Verb-Cot* and *Verb-Topk* enhanced metacognitive ability through chain-of-thought reasoning (Wei et al. 2022) and prompting candidate answers from the LLM, respectively. More specifically, *Verb-Topk* demonstrated notable improvement, likely due to its effective mitigation of initial overconfidence issues. Despite these advancements, *Verb-Cot* and *Verb-Topk* still grapple with overconfidence, resulting in lower metacognitive ability compared to consistency-based strategies. In consistency methods, *Perturbing* and *Misleading* achieve metacognitive ability slightly lower than that of *Self-Random*, possibly due to the introduction of irrelevant reasoning paths by additional perturbations. What's more, by defining metacognitive ability as the capacity to evaluate response correctness through confidence, our DMC provides a task-agnostic approach to assess the effectiveness of confidence elicitation methods.

Results and Analysis of RQ3

As shown in Table 3, we selected three representative confidence elicitation methods to assess the metacognitive abilities quantified by the DMC framework across different large models (including LLaMA2, GPT-3.5, and GPT-4) on all datasets. Experimental results indicate that regardless of the confidence elicitation method used, GPT-4 outperforms GPT-3.5, which in turn outperforms LLaMA2 (with average metacognitive abilities per LLM across three C.E.Ms being 0.5491 vs. 0.4835 vs. 0.3210, respectively). Therefore, it is apparent that the performance of large models is positively linked to their displayed metacognitive ability. This connection may be attributed to the complexity of the models, with more intricate structures and parameters enabling a better capture of language complexity and contextual information, leading to enhanced metacognitive performance. Additionally, it could be due to the advanced models' inclination to

C.E.M	SAT	HSM	CQA	GFacts	Bool	Date	Med	Ethics	Average
Verb-Vanilla	0.3192	0.3276	0.3203	0.3215	0.3188	0.3236	0.3271	0.3227	0.3225
Verb-Cot	0.3739	0.3770	0.3785	0.3717	0.3741	0.3769	0.3804	0.3820	0.3768
Verb-Topk	0.5288	0.5325	0.5344	0.5278	0.5312	0.5273	0.5309	0.5225	0.5294
Self Random	0.6059	0.5976	0.5938	0.5981	0.6027	0.5944	0.5955	0.5956	0.5986
Perturbing	0.5599	0.5596	0.5585	0.5641	0.5615	0.5633	0.5584	0.5542	0.5599
Misleading	0.5600	0.5652	0.5619	0.5611	0.5547	0.5573	0.5585	0.5634	0.5606

Table 2: Comparison of metacognitive ability quantified using the DMC Framework with different confidence elicitation methods across all datasets.

Model	C.E.M	SAT	HSM	CQA	GFacts	Bool	Date	Med	Ethics	Average
LLaMA2	Verb-Vanilla	0.2485	0.2498	0.2476	0.2507	0.2525	0.2557	0.2448	0.2546	0.2505
	Verb-Topk	0.3293	0.3359	0.3348	0.3273	0.3380	0.3276	0.3365	0.3347	0.3330
	Self Random	0.3832	0.3788	0.3817	0.3837	0.3779	0.3753	0.3765	0.3801	0.3797
GPT-3.5	Verb-Vanilla	0.3192	0.3276	0.3203	0.3215	0.3188	0.3236	0.3271	0.3227	0.3225
	Verb-Topk	0.5288	0.5325	0.5344	0.5278	0.5312	0.5273	0.5309	0.5225	0.5294
	Self Random	0.6059	0.5976	0.5938	0.5981	0.6027	0.5994	0.5955	0.5956	0.5986
GPT-4	Verb-Vanilla	0.3576	0.3590	0.3655	0.3663	0.3679	0.3653	0.3614	0.3659	0.3636
	Verb-Topk	0.5568	0.5632	0.5593	0.5641	0.5626	0.5618	0.5646	0.5572	0.5612
	Self Random	0.7195	0.7205	0.7227	0.7273	0.7184	0.7267	0.7244	0.7215	0.7226

Table 3: Comparsion of metacognitive ability quantified using the DMC Framework across various LLMs.

mimic human language understanding and expression, including confidence representation. Simultaneously, different LLMs exhibit varying sensitivities to confidence elicitation methods, potentially resulting in diverse metacognitive performances. As depicted in Table 3, LLaMA2 demonstrates limited sensitivity to different C.E.Ms (max=0.3797, min=0.2505, st.d.=0.0534). Conversely, GPT-4 shows larger variations in metacognitive abilities when employing different C.E.Ms (max=0.7226, min=0.3636, st.d.=0.1468).

Results and Analysis of RQ4

We utilized the DMC framework to evaluate hallucination mitigation in LLMs via the AbstainQA task (Feng et al. 2024). Specifically, Reliable Accuracy (R-Acc) and Abstain Accuracy (A-Acc) were employed as metrics to measure task performance, with higher values indicating better performance. In Figure 4, the x-axis illustrates the combinations of three LLMs and three C.E.Ms, with different colors (orange, blue, and green) representing the performance of each combination in metacognition, R-Acc, and A-Acc on the y-axis. The three distinct colored lines exhibit similar trends, suggesting that higher metacognitive ability correspond to better performance in the AbstainQA task, underscoring the potential of enhancing metacognition in LLMs to alleviate hallucination issues.

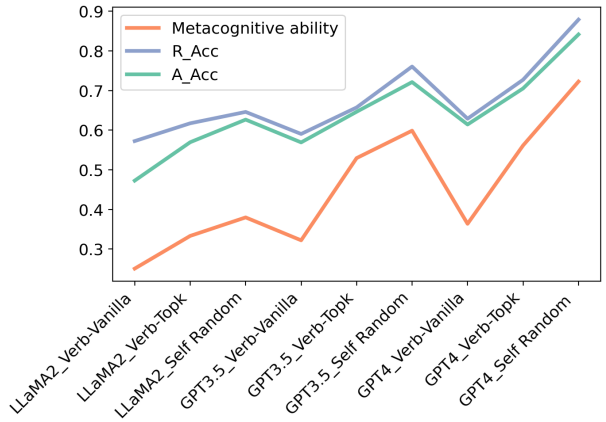

Figure 4: LLMs' metacognitive ability and their performance on the AbstainQA task.

Conclusion

In this paper, we introduce a novel general framework for quantifying metacognitive ability in large language models, named DMC. Through comprehensive experiments and analyses, we have observed that DMC successfully separates metacognitive from cognitive capabilities, enabling a more precise quantification of metacognitive ability. In addition, the quantification of metacognitive ability is influenced by the choice of confidence elicitation methods and varies across different large language models. Moreover, improving the metacognitive ability of LLMs shows potential in mitigating hallucination issues.

Acknowledgements

This work is funded by National Natural Science Foundation of China (under project No. 62377013), Natural Science Foundation of Shanghai, China (under project No. 22ZR1419000), and the Fundamental Research Funds for the Central Universities. It is also supported by STI 2030-Major Projects 2021ZD0200500, the Research Project of Changning District Science and Technology Committee (under project No. CNKW2022Y37), and the Medical Master's and Doctoral Innovation Talent Base Project of Changning District (under project No. RCJD2022S07).

References

Abdi, H. 2010. Coefficient of variation. *Encyclopedia of Research Design*, 1(5): 169–171.

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*.

Brier, G. W. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1): 1–3.

Chen, J.; Lin, H.; Han, X.; and Sun, L. 2024a. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 17754–17762.

Chen, K.; Chen, Q.; Zhou, J.; Yishen, H.; and He, L. 2024b. DiaHalu: A Dialogue-level Hallucination Evaluation Benchmark for Large Language Models. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 9057–9079. Miami, Florida, USA: Association for Computational Linguistics.

De Martino, B.; Fleming, S. M.; Garrett, N.; and Dolan, R. J. 2013. Confidence in value-based choice. *Nature Neuroscience*, 16(1): 105–110.

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

Feng, S.; Shi, W.; Wang, Y.; Ding, W.; Balachandran, V.; and Tsvetkov, Y. 2024. Don't Hallucinate, Abstain: Identifying LLM Knowledge Gaps via Multi-LLM Collaboration. *arXiv preprint arXiv:2402.00367*.

Fleming, S. M. 2024. Metacognition and confidence: A review and synthesis. *Annual Review of Psychology*, 75(1): 241–268.

Fleming, S. M.; and Dolan, R. J. 2012. The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594): 1338–1349.

Fleming, S. M.; and Lau, H. C. 2014. How to measure metacognition. *Frontiers in Human Neuroscience*, 8: 443.

Gekhman, Z.; Yona, G.; Aharoni, R.; Eyal, M.; Feder, A.; Reichart, R.; and Herzig, J. 2024. Does Fine-Tuning LLMs on New Knowledge Encourage Hallucinations? *arXiv preprint arXiv:2405.05904*.

Green, D. M.; Swets, J. A.; et al. 1966. *Signal detection theory and psychophysics*, volume 1. Wiley New York.

Guan, X.; Liu, Y.; Lin, H.; Lu, Y.; He, B.; Han, X.; and Sun, L. 2024. Mitigating large language model hallucinations via autonomous knowledge graph-based retrofitting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18126–18134.

Gunjal, A.; Yin, J.; and Bas, E. 2024. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18135–18143.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *International conference on Machine Learning*, 1321–1330. PMLR.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2020. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*.

Kadavath, S.; Conerly, T.; Askell, A.; Henighan, T.; Drain, D.; Perez, E.; Schiefer, N.; Hatfield-Dodds, Z.; DasSarma, N.; Tran-Johnson, E.; et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Li, H.; Wu, W.; Ji, Y.; Zheng, H.; Shi, L.; Chen, X.; and He, L. 2024. MetaESC: Enhancing Emotional Support Conversation through Metacognition. In *International Conference on Database Systems for Advanced Applications*, 337–348. Springer.

Lin, Z.; Trivedi, S.; and Sun, J. 2023. Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models. *Transactions on Machine Learning Research*, 2024.

Maniscalco, B.; and Lau, H. 2012. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, 21(1): 422–430.

Mason, I. B. 2003. Binary events. Forecast Verification: A Practitioner's Guide in Atmospheric Science, IT Jolliffe and DB Stephenson, Eds.

Masson, M. E.; and Rotello, C. M. 2009. Sources of bias in the Goodman–Kruskal gamma coefficient measure of association: Implications for studies of metacognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(2): 509.

Mielke, S. J.; Szlam, A.; Boureau, Y.-L.; and Dinan, E. 2020. Linguistic calibration through metacognition: aligning dialogue agent responses with expected correctness. *arXiv preprint arXiv:2012.14983*, 11.

OpenAI. 2021. ChatGPT. https://www.openai.com/gpt-3/. Accessed: April 21, 2023.

Suzgun, M.; Scales, N.; Schärli, N.; Gehrmann, S.; Tay, Y.; Chung, H. W.; Chowdhery, A.; Le, Q.; Chi, E.; Zhou, D.; et al. 2023. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. In *Findings of the Association for Computational Linguistics: ACL*, 13003–13051.

Talmor, A.; Herzig, J.; Lourie, N.; and Berant, J. 2019. CommonsenseQA: A Question Answering Challenge Targeting

Commonsense Knowledge. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4149–4158.

Tian, K.; Mitchell, E.; Zhou, A.; Sharma, A.; Rafailov, R.; Yao, H.; Finn, C.; and Manning, C. D. 2023. Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 5433–5442.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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.

Xiong, M.; Hu, Z.; Lu, X.; LI, Y.; Fu, J.; He, J.; and Hooi, B. 2024. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. In *In International Conference on Learning Representations*.

Xiong, M.; Li, S.; Feng, W.; Deng, A.; Zhang, J.; and Hooi, B. 2022. Birds of a feather trust together: Knowing when to trust a classifier via adaptive neighborhood aggregation. *arXiv preprint arXiv:2211.16466*.

Yu, J.; Wang, X.; Tu, S.; Cao, S.; Zhang-Li, D.; Lv, X.; Peng, H.; Yao, Z.; Zhang, X.; Li, H.; et al. 2023. KoLA: Carefully Benchmarking World Knowledge of Large Language Models. In *In International Conference on Learning Representations*.

Zhong, W.; Cui, R.; Guo, Y.; Liang, Y.; Lu, S.; Wang, Y.; Saied, A.; Chen, W.; and Duan, N. 2024. AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, 2299–2314.

Zhou, K.; Jurafsky, D.; and Hashimoto, T. 2023. Navigating the Grey Area: How Expressions of Uncertainty and Overconfidence Affect Language Models. In *The Conference on Empirical Methods in Natural Language Processing*.

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*, 1453–1463.