Decoupling Knowledge and Reasoning in LLMs: An Exploration Using Cognitive Dual-System Theory

Mutian Yang[1], **Jiandong Gao**[1, *] **Ji Wu**[1, 2, 3, *]

[1]Department of Electronic Engineering, Tsinghua University, Haidian District, Beijing, 100084
[2]Beijing National Research Center for Information Science and Technology, Haidian District, Beijing, 100084
[3]College of AI, Tsinghua University, Haidian District, Beijing, 100084
{yangmutian, jdgao, wuji_ee}@tsinghua.edu.cn

Abstract

While large language models (LLMs) leverage both knowledge and reasoning during inference, the capacity to distinguish between them plays a pivotal role in model analysis, interpretability, and development. Inspired by dual-system cognitive theory, we propose a cognition attribution framework to decouple the contribution of knowledge and reasoning. In particular, the cognition of LLMs is decomposed into two distinct yet complementary phases: knowledge retrieval (Phase 1) and reasoning adjustment (Phase 2). To separate these phases, LLMs are prompted to generate answers under two different cognitive modes, fast thinking and slow thinking, respectively. The performance under different cognitive modes is analyzed to quantify the contribution of knowledge and reasoning. This architecture is employed to 15 LLMs across 3 datasets. Results reveal: (1) reasoning adjustment is domain-specific, benefiting reasoning-intensive domains (e.g., mathematics, physics, and chemistry) and potentially impairing knowledge-intensive domains. (2) Parameter scaling improves both knowledge and reasoning, with knowledge improvements being more pronounced. Additionally, parameter scaling make LLMs reasoning significantly more prudent, while moderately more intelligent. (3) Knowledge primarily resides in lower network layers, while reasoning operates in higher layers. Our framework not only helps understand LLMs from a "decoupling" perspective, but also provides new insights into existing research, including scaling laws, hierarchical knowledge editing, and limitations of small-scale-LLM reasoning.

Code — https://github.com/yangmutian/decoupling

Introduction

LLMs have garnered significant research attention due to remarkable capabilities (Achiam et al. 2023; Yang et al. 2024a; Touvron et al. 2023). Building upon the foundation of scaling laws (Kaplan et al. 2020), various methodologies (including pretraining (Devlin et al. 2019), continual pretraining (Ke et al. 2023; Yang et al. 2024b; Xie, Aggarwal, and Ahmad 2024), and instruction tuning (Hu et al. 2022; Li and Liang 2021; Liu et al. 2021)) are developed to inject knowledge into LLMs. However, their reasoning capability remains disputed, as they often struggle with complex reasoning tasks (Mirzadeh et al. 2024; Huang et al. 2025a).

To address this challenge, Chain-of-Thought (CoT) is proposed, enabling LLMs to mimic human-like progressive reasoning by generating intermediate reasoning steps (Wei et al. 2022). However, early approaches to CoT generation typically relied on domain-specific prompt engineering, lacking the capability to automatically produce universally applicable reasoning chains across diverse domains.

The emergence of reasoning LLMs, such as OpenAI o1, enables automatic generation of universal CoT through distillation and reinforcement learning (OpenAI 2023). Although the specifics of o1 remain undisclosed, extensive replication efforts have successfully produced LLMs with powerful reasoning capability (Qin et al. 2024; Huang et al. 2024, 2025b). The breakthrough demonstrates that LLMs possess not only extensive knowledge but also advanced reasoning abilities.

In this context, it is scientifically important to distinguish between the contribution of knowledge and reasoning, as this is crucial for understanding the inference behaviours of LLMs. However, the joint employment of knowledge and reasoning during inference make it hard to discern their contribution.

For this purpose, we propose a cognition attribution framework based on dual-system cognitive theory, which decomposes LLMs inference into two distinct but complementary phases: (1) knowledge retrieval (Phase 1), where LLMs rapidly generate initial responses by accessing learned information, and (2) reasoning adjustment (Phase 2), where they refines the initial responses through CoT generation. To separate the two cognitive phases, LLMs are prompted to generate answers under two distinct cognitive modes: fast thinking and slow thinking. During fast thinking, LLMs experience Phase 1, while during slow thinking, LLMs rely on both Phase 1 and Phase 2. The difference between cognitive modes is analyzed to decouple knowledge and reasoning. Our main findings include:

- The contribution of reasoning adjustment varies across domains. It plays more crucial roles in some reasoning-intensive domains (such as mathematics, physics, chemistry) than the others.
- Parameter scaling enhances both knowledge and reasoning, with knowledge being the dominant factor. Addi-

* Corresponding Authors

tionally, parameter scaling makes the reasoning significantly more ”prudent” in all domains and moderately more ”intelligent” in some specific domains.

• Knowledge retrieval primarily occurs in lower network layers, while reasoning adjustments are localized in higher layers, suggesting a functional separation in cognition.

In conclusion, our study presents a cognition attribution architecture that decouples knowledge and reasoning in LLMs. This framework not only offers a novel perspective on the cognitive characteristics of LLMs, but also provides new insights into related areas of research, including scaling laws (Kaplan et al. 2020), hierarchical knowledge editing (Zhang, Li, and Wu 2024; Meng et al. 2022a,b), and limitations of small-model reasoning (Li et al. 2025).

From Dual-System Cognitive Theory to Cognition Attribution Architecture

LLMs perform inference through the integrated application of knowledge and reasoning, making it challenging to isolate and evaluate their respective contributions. To address this issue, we propose a cognition attribution architecture based on dual-system cognitive theory, as shown in Figure 1. This architecture decouples knowledge and reasoning in four steps.

Step 1: Define the Respective Roles of Knowledge and Reasoning

According to dual-system cognitive theory (Kahneman 2011), humans exhibit two distinct yet complementary cognitive phases: knowledge retrieval (Phase 1) and reasoning adjustment (Phase 2). This study maps these concepts onto LLMs, hypothesizing LLMs also generate answers in two phases. In Phase 1, LLMs retrieves memorized knowledge purely based on the input. In Phase 2, LLMs applied CoT reasoning to adjust these initial retrievals.

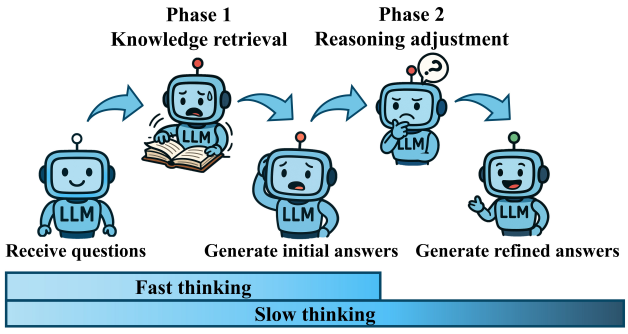

Figure 1: The schema of the cognition attribution architecture, supposing the cognitive process of LLMs involves two primary phases: knowledge retrieval (Phase 1) and reasoning adjustment (Phase 2). In the fast thinking mode, LLMs rely solely on knowledge retrieval. In contrast, in the slow thinking mode, they leverage both knowledge retrieval and reasoning adjustment.

Fast Thinking

Your only task is to select the most appropriate answer from the given options without any reasoning.
Output format: A single letter representing your answer (A, B, C, D)
Output rules:
- Output EXACTLY one letter (A, B, C, D)
- Do not provide any explanation or reasoning
- Do not restate the question or options
- Do not include any additional text or punctuation

Figure 2: Fast thinking prompt

Slow Thinking

Your role as an assistant involves thoroughly exploring questions through a systematic long thinking process before providing the final precise and accurate solutions. This requires engaging in a comprehensive cycle of analysis, summarizing, exploration, reassessment, reflection, backtracing, and iteration to develop well-considered thinking process.
Please structure your response into two main sections: Thought and Solution.
In the Thought section, detail your reasoning process. Each step should include detailed considerations such as analisying questions, summarizing relevant findings, brainstorming new ideas, verifying the accuracy of the current steps, refining any errors, and revisiting previous steps.
In the Solution section, based on various attempts, explorations, and reflections from the Thought section, systematically present the final solution that you deem correct. The solution should remain a logical, accurate, concise expression style and detail necessary step needed to reach the conclusion.
Now, try to solve the following question through the above guidelines:

Figure 3: Slow thinking prompt

Step 2: Separate Knowledge Retrieval and Reasoning Adjustment

To separate these phases, we further introduce the concept of cognitive modes in dual-system cognitive theory - fast thinking and slow thinking. LLMs are prompted to generate responses under each modes separately. The prompt guiding LLMs to conduct fast and slow thinking are:

During fast thinking, LLMs generates responses y_{fast} to query Q directly based on the input question x without any additionally reasoning. In this mode, LLMs rely solely on Phase 1, with its performance depends entirely on the knowledge retrieval capability, denoted as C_{KR}. The process is formally defined as:

$$y_{\text{fast}} = \arg\max P(y \mid x = \text{Q}; C_{\text{KR}}) \quad (1)$$

During slow thinking, LLMs first produces initial answers y_{fast} similar to fast thinking, and then adjusts initial answers through CoT reasoning to get refined answers y_{slow}. In this mode, LLMs experience both Phase 1 and Phase 2. The performance relies on both knowledge retrieval capability C_{KR} and reasoning adjustment capability C_{RA}, defined as:

$$y_{\text{slow}} = \arg\max P(y \mid y_{\text{fast}}, x = \text{Q}; C_{\text{KR}}, C_{\text{RA}}) \quad (2)$$

Step 3: Determine the Contribution of Knowledge Retrieval and Reasoning Adjustment

The accuracy of fast and slow thinking is evaluated to measure the retrieval capability C_{KR} and reasoning adjustment

capability C_{RA}:

$$C_{KR} := A_{\text{fast}} = \frac{1}{|D|} \sum_{x \in D} I(y_{\text{fast}} = y^*) \tag{3}$$

$$C_{KR} + C_{RA} := A_{\text{slow}} = \frac{1}{|D|} \sum_{x \in D} I(y_{\text{slow}} = y^*) \tag{4}$$

A_{fast} and A_{slow} denote the accuracy of fast and slow thinking on dataset D, y^* denotes the true answer, and $|D|$ indicates the total number of problems. The indicator function $I(\cdot)$ returns 1 if the condition is satisfied, and 0 otherwise. By subtracting Equation 3 from Equation 4, we obtain:

$$\begin{aligned} C_{RA} := \delta &= A_{\text{slow}} - A_{\text{fast}} \\ &= \frac{1}{|D|}\left(\sum_{x \in D} I(y_{\text{slow}} = y^*) - \sum_{x \in D} I(y_{\text{fast}} = y^*)\right) \end{aligned} \tag{5}$$

This equation shows that the accuracy difference between slow and fast thinking, referred to as the reasoning gain δ, reflects the capability of reasoning adjustment C_{RA}. Ultimately, we decouple the knowledge and reasoning capabilities in Equation 3 and Equation 5.

Step 4: Decompose Reasoning Adjustment Into Correction and Overthinking

Based on the properties of the indicator function:

$$\begin{aligned} \sum_{x \in D} I(y_{\text{slow}} = y^*) = \sum_{x \in D} I(y_{\text{slow}} = y^*) \cdot \\ (I(y_{\text{fast}} = y^*) + I(y_{\text{fast}} \neq y^*)) \end{aligned} \tag{6}$$

$$\begin{aligned} \sum_{x \in D} I(y_{\text{fast}} = y^*) = \sum_{x \in D} I(y_{\text{fast}} = y^*) \cdot \\ (I(y_{\text{slow}} = y^*) + I(y_{\text{slow}} \neq y^*)) \end{aligned} \tag{7}$$

δ is derived as

$$\begin{aligned} \delta = &\underbrace{\frac{1}{|D|} \sum_{x \in D} I(y_{\text{fast}} \neq y^*) \cdot I(y_{\text{slow}} = y^*) -}_{\delta_c} \\ &\underbrace{\frac{1}{|D|} \sum_{x \in D} I(y_{\text{fast}} = y^*) \cdot I(y_{\text{slow}} \neq y^*)}_{\delta_o} \end{aligned} \tag{8}$$

This shows reasoning adjustment has two effects: it corrects the errors made by fast thinking, but also incorrectly override the correct answers. We refer to these effects as correction and overthinking, denoting their corresponding accuracy changes as δ_c and δ_o. Their formal definitions are derived as:

$$\begin{aligned} \delta_c = &\frac{\sum_{x \in D} I(y_{\text{fast}} \neq y^*) \cdot I(y_{\text{slow}} = y^*)}{\sum_{x \in D} I(y_{\text{fast}} \neq y^*)} \cdot \\ &\sum_{x \in D} I(y_{\text{fast}} \neq y^*) = r_c \cdot |D_{fast}^{false}| \end{aligned} \tag{9}$$

$$\begin{aligned} \delta_o = &\frac{\sum_{x \in D} I(y_{\text{fast}} = y^*) \cdot I(y_{\text{slow}} \neq y^*)}{\sum_{x \in D} I(y_{\text{fast}} = y^*)} \cdot \\ &\sum_{x \in D} I(y_{\text{fast}} = y^*) = r_o \cdot |D_{fast}^{true}| \end{aligned} \tag{10}$$

$|D_{fast}^{false}|$ and $|D_{fast}^{true}|$ denote the numbers of incorrect and correct answers produced by fast thinking, respectively. r_c and r_o represent the rates at which the LLMs correct wrong answers and overthink correct answers. Thus, the reasoning gain δ is expressed as:

$$\delta = \delta_c - \delta o = \frac{1}{|D|}\left(r_c \cdot |D_{fast}^{false}| - r_o \cdot |D_{fast}^{true}|\right) \tag{11}$$

δ arises from two opposing components: the benefit from correction δ_c and the loss from overthinking δ_o. $|D_{\text{fast}}^{\text{false}}|$ and $|D_{\text{fast}}^{\text{true}}|$, determined by knowledge retrieval, define the upper bounds for correction and overthinking. δ_c and δ_o, influenced by reasoning adjustment, define the actual levels of correction and overthinking in the candidates.

Some research suggests reasoning benefits LLMs by providing information gain (Ton, Taufiq, and Liu 2024), while others attribute reasoning failure to noise introduction (Gan, Liao, and Liu 2025). Equation 11 integrates these insights, indicating it is the trade off between information gain and noise that determines the effect of reasoning.

Results

Experiment Procedure

Dataset MMLU, MathQA, and MedQA are employed (Hendrycks et al. 2020; Zhang et al. 2018; Amini et al. 2019). MMLU serves as a general dataset, while MathQA and MedQA are domain-specific datasets. Although each MMLU question is annotated with a subject, the 57-in-total subjects are too fine-grained for our analysis. Therefore, we group them into 13 broader domains, as shown as follow:

Mathematics: Abstract Algebra, College Mathematics, Elementary Mathematics, High School Mathematics, High School Statistics.

Physics: College Physics, Conceptual Physics, High School Physics.

Chemistry: College Chemistry, High School Chemistry.

Computer Science: College Computer Science, High School Computer Science, Machine Learning.

Economics and Business: Business Ethics, Econometrics, High School Macroeconomics, High School Microeconomics, Management, Marketing, Professional Accounting.

Biology and Medicine: Anatomy, Clinical Knowledge, College Biology, College Medicine, High School Biology,

Human Sexuality, Medical Genetics, Nutrition, Professional Medicine, Virology.

Psychology and Sociology: High School Psychology, Human Aging, Professional Psychology, Sociology.

Geography and Astronomy: Astronomy, Global Facts, High School Geography, World Religions.

Engineering: Computer Security, Electrical Engineering.

Philosophy: Formal Logic, Logical Fallacies, Moral Disputes, Moral Scenarios, Philosophy.

Law: International Law, Jurisprudence, Professional Law.

History: High School European History, High School US History, High School World History, Prehistory.

Domain 13 (Political Science): High School Government and Politics, Public Relations, Security Studies, US Foreign Policy.

To ensure comparability between datasets, the four-option version of MedQA is selected. For MathQA, one wrong option is randomly removed to reduce the choices from five to four.

Model To enhance the generalizability of our work, extensive experiments are conducted on 15 LLMs, including *Qwen*, *LLaMA*, *Gemma*, *Phi*, and *GLM* (Yang et al. 2024a; Grattafiori et al. 2024; Team et al. 2024; Abdin et al. 2024; GLM et al. 2024).

Small LLMs Overthink More Than Correct During Reasoning

Although LLMs have been extensively studied on MMLU, most existing studies fail to disentangle the contributions of knowledge retrieval and reasoning adjustment. Our cognition attribution architecture is employed to decouple their contribution on MMLU.

Table 1 reports the accuracy of fast and slow thinking in "Fast" and "Slow" columns. Reasoning gain δ is determined by the accuracy difference between fast and slow thinking, and subsequently decomposed into δ_c and δ_o according to Equation 11.

Reasoning adjustment benefits most models, yielding positive δ. However, for some extremely small models (marked in underline), δ becomes zero or even negative, as the loss from overthinking outweighs the benefit from correction.

The correction rate r_c and overthinking rate r_o are investigated. Small LLMs exhibit lower r_c and higher r_o compared to their larger counterparts. Moreover, the variation in r_o across model sizes is more prominent than that in r_o. Notably, the r_o of *LLaMA 1B* is 45.4% higher than that of *LLaMA 70B*, while its r_c is merely 8.7% lower. This asymmetry suggests that the substantial overthinking tendency, rather than the modest correction capability, plays a greater role in the negative reasoning gain in small models. More discussion on correction and overthinking is in Section .

Our results offer new insights into prior findings. (Liu et al. 2024; Cuadron et al. 2025) report that reasoning negatively affects performance, while (Gan, Liao, and Liu 2025) attributes this to noise introduction. Our study further reveals that small LLMs are prone to introducing noise and leading to overthinking. (Li et al. 2025) finds that small models benefit less from long CoT distillation. Our results suggest that is because small models introduce more noise when CoT becomes longer. In summary, our study indicates that small LLMs tend to introduce noise during reasoning, reflecting a lack of "prudence" during inference.

Model	**Fast**	**Slow**	δ	δ_c	δ_o	r_c	r_o
Qwen 1.5B	53.9	50.0	-3.9	13.8	17.7	29.9	32.8
Qwen 3B	60.9	63.0	2.1	14.1	12.0	35.9	19.7
Qwen 7B	67.8	71.8	4.0	11.8	7.7	36.5	11.4
Qwen 14B	75.2	78.7	3.6	9.1	5.5	36.5	7.3
Qwen 32B	79.5	81.1	1.5	6.9	5.4	33.9	6.8
QwQ 32B	77.7	85.6	7.9	10.0	10.0	54.6	5.5
LLaMA 1B	35.1	35.1	0.1	17.9	17.3	29.4	52.6
LLaMA 3B	52.4	55.9	3.5	17.1	13.6	35.9	25.9
LLaMA 8B	60.3	65.7	5.4	15.9	10.4	40.0	17.3
LLaMA 70B	81.1	82.4	1.3	7.2	5.9	38.1	7.2
Gemma 2B	53.5	50.9	-2.6	10.4	13.0	22.5	24.4
Gemma 9B	69.3	69.5	0.3	9.5	9.2	30.8	13.3
Gemma 27B	72.5	73.5	1.0	9.4	8.4	34.2	11.6
GLM 9B	63.8	70.1	6.3	14.5	8.2	40.0	13.0
Phi 14B	78.1	84.1	6.0	10.2	4.2	46.7	5.4

Table 1: Performance on MMLU using cognition attribution architecture. "Fast" and "Slow" columns represent the accuracy under fast and slow thinking, respectively.

Reasoning Exhibits Significantly Domain-Specific Variability

While it is widely hypothesized that reasoning adjustment is domain-specific (e.g., mathematics is regarded as reasoning-intensive), the claim remains imprecise and speculative. We employ cognition attribution architecture to quantify the contribution of reasoning adjustment across domains.

Table 2 reports the reasoning gain δ for 15 LLMs across 13 domains, spanning natural sciences to humanities. The results reveal substantial cross-domain variation in δ. Notably, *Qwen 1.5B* shows an inter-domain accuracy gap of 34.7% (22.3% for Mathematics and -12.4% for Political Science).

While domains exhibit varying sensitivity to reasoning adjustment, their relative ranking remains stable. The top-1, top-2, and top-3 δ for each model are highlighted in **underline&bold**, underline, and **bold**, respectively. Across 15 models, δ of mathematics, physics, and chemistry ranks among the top-3 for 15, 14, and 12 times, respectively, indicating that these domains consistently benefit the most from reasoning adjustment. In contrast, political science and history yield negative δ in 10 and 9 models, respectively, suggesting that reasoning adjustment even impair their performance.

Our study provides evidence that reasoning adjustment exhibits significant domain-specific variability. This phenomenon is explained according to the relationship between

Model	Mat	Phy	Che	CS	Eco Bus	Phi	Geo Ast	Bio Med	Psy Soc	Eng	Law	His	Pol
Qwen 1.5B	**22.3**	7.2	**5.6**	-2.2	-4.1	-7.4	-5.6	-6.8	-7.5	-6.9	-6.9	-9.0	-12.4
Qwen 3B	**25.8**	15.4	**10.9**	2.9	0.3	8.3	2.7	-4.7	-2.3	-3.3	0.8	-3.9	-7.1
Qwen 7B	**27.6**	12.1	**10.6**	8.3	3.2	10.3	-0.2	-0.6	-0.6	-0.8	0.2	-1.8	-1.9
Qwen 14B	**20.0**	**12.3**	13.5	7.4	2.3	2.3	1.5	1.8	1.7	4.5	0.9	-1.0	-1.9
Qwen 32B	**13.9**	8.8	**6.6**	6.4	2.7	3.8	-0.6	-1.1	-1.1	-1.2	-2.3	-2.0	0.5
QwQ 32B	**22.7**	**16.9**	12.6	11.5	6.1	19.7	2.3	4.6	3.4	6.7	10.5	4.1	-1.5
LLaMA 1B	**5.9**	**7.3**	-0.8	-5.2	4.0	-2.1	4.7	4.3	6.3	-0.2	-7.7	-2.0	-3.8
LLaMA 3B	**20.6**	7.6	**8.3**	-1.6	2.4	0.6	13.4	5.9	0.2	-3.7	-2.7	1.1	1.2
LLaMA 8B	**22.5**	15.0	**10.6**	5.8	4.7	1.9	4.9	4.0	4.9	-0.4	0.6	0.7	2.8
LLaMA 70B	**16.4**	**7.0**	9.2	6.7	3.1	-0.1	-2.0	-0.4	-0.2	-2.5	-2.4	-0.7	-0.3
Gemma 2B	4.9	**5.7**	**-0.7**	-3.3	-3.6	-7.7	-5.7	-2.3	-2.9	-6.9	-3.6	-4.2	-10.0
Gemma 9B	**12.7**	10.0	**8.9**	2.9	-0.2	-3.5	-2.0	-2.2	-2.1	-2.9	-3.0	-1.9	-5.7
Gemma 27B	**19.9**	10.3	5.0	**5.2**	0.7	-4.7	-2.9	-2.9	-3.1	-0.4	-0.5	-0.7	-6.8
GLM 9B	**24.5**	16.8	**16.2**	11.2	5.4	3.7	3.7	4.0	2.5	9.4	-0.8	-0.5	1.4
Phi 14B	**25.9**	19.5	**12.2**	10.6	6.9	2.3	1.7	3.3	1.0	5.3	3.8	1.5	0.2
Average	**19.0**	11.5	**8.6**	4.4	2.3	1.8	1.1	0.5	0.0	-0.2	-0.9	-1.4	-3.0

Table 2: δ of 15 LLMs across 13 domains on MMLU. For each model, the top-1, top-2, and top-3 δ are highlighted in **underline&bold**, underline, and **bold**, respectively. Mat, Mathematics; Phy, Physics; Che, Chemistry; CS, Computer Science; EcoBus, Economics and Business; Phi: Philosophy; GeoAst: Geography and Astronomy; BioMed, Biology and Medicine; PhySoc: Psychology and Sociology; Eng, Engineering; His: History; Pol: Political Science.

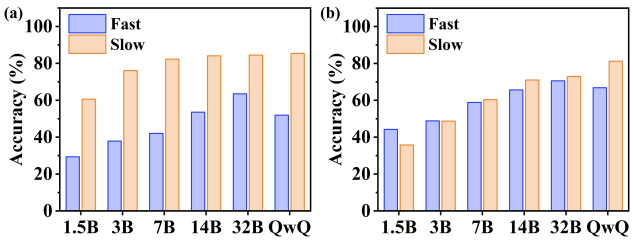

Figure 4: The accuracy of the *Qwen* series on (a) MathQA and (b) MedQA. The results demonstrate that mathematics imposes higher demands on reasoning than medicine.

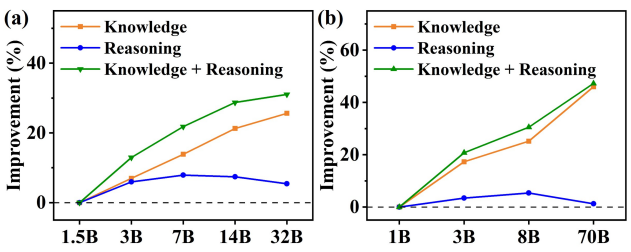

Figure 5: Improvement of knowledge retrieval and reasoning adjustment with parameter scaling for *Qwen* series. The plotted values quantify improvement relative to the smallest baseline model in each series: *Qwen 1.5B* for the *Qwen* series in panel (a) and *LLaMA 1B* for the *LLaMA* series in panel (b). The green, orange, and blue trajectories represent the $C_{KR}+C_{RA}$, C_{KR}, and C_{RA} according to Equation 3-5.

correction and overthinking in Equation 11. In knowledge-intensive domains, LLMs make mistakes due to insufficient knowledge. Reasoning cannot compensate for the knowledge gap and instead introduces additional noise. Consequently, δ_c falls short of δ_o, yielding negative δ in these domains. In reasoning-intensive domains (notably mathematics, physics, and chemistry), mistakes often arise from generating answers under high uncertainty. Reasoning effectively reduce ambiguity and correct mistakes. Therefore, δ_c outweighs δ_o, resulting in positive δ in these domains.

In addition to MMLU, this paper also study the performance of *Qwen* series LLMs on MathQA and MedQA. The results on MathQA and MedQA show high consistency with MMLU (Figure 4). A significantly high δ (with a range from 20.9% to 40.2%) is observed on MathQA, while the δ on MedQA ranges sorely from -8.5% to 14.3%.

The results reveal that although both knowledge retrieval and reasoning adjustment benefit from model scaling, the improvement in knowledge retrieval is more pronounced and sustained. Reasoning adjustment shows notable gains primarily when scaling from small-sized models (1B) to medium-sized models (8B), but exhibits limited improvement beyond this range. Table 1 demonstrates the observed enhancement in reasoning adjustment from small to medium size is largely attributed to a reduction in overthinking. Once the model reaches medium size, it becomes sufficiently "prudent" to avoid overthinking. Consequently, further scaling leads to only marginal reductions in overthinking, re-

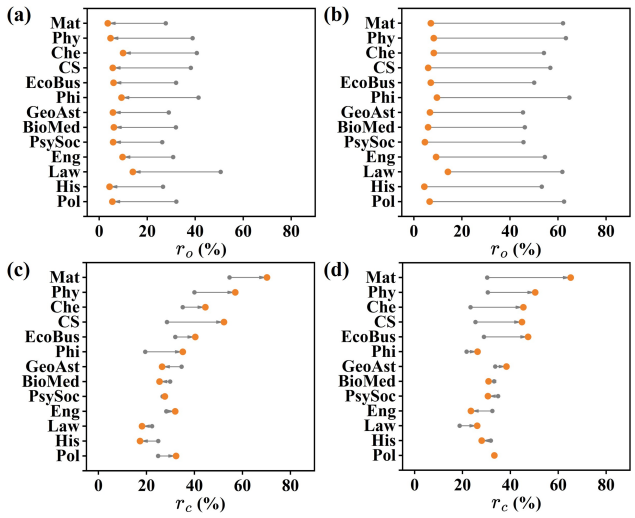

Figure 6: Impact of parameter scaling on overthinking rate r_o and correction rate r_c. (a) and (b) show the r_o of *Qwen* (orange for 32B and grey for 1.5B) and *LLaMA* (orange for 70B and grey for 1B), respectively. (c) and (d) show the r_c of *Qwen* (orange for 32B and grey for 1.5B) and *LLaMA* (orange for 70B and grey for 1B), respectively. Mat, Mathematics; Phy, Physics; Che, Chemistry; CS, Computer Science; EcoBus, Economics and Business; Phi: Philosophy; GeoAst: Geography and Astronomy; BioMed, Biology and Medicine; PhySoc: Psychology and Sociology; Eng, Engineering; His: History; Pol: Political Science.

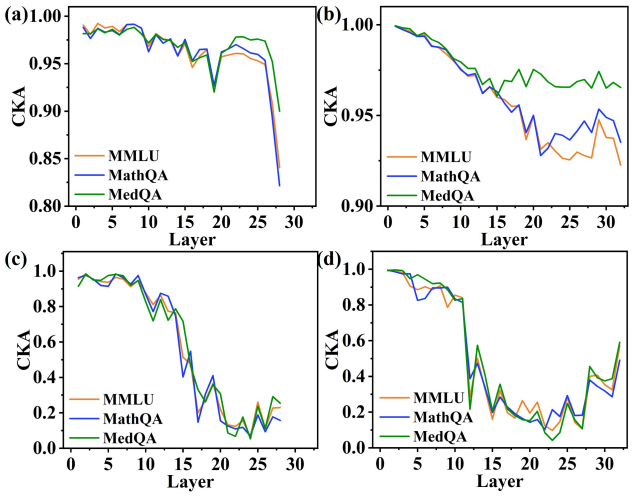

Figure 7: CKA between corresponding layers under slow and fast thinking. Lower CKA indicates reduced similarity for the given layer. (a, b) Output similarity for *Qwen 7B* and *LLaMA 8B*. (c, d) Attention similarity for *Qwen 7B* and *LLaMA 8B*.

sulting in limited additional gains in reasoning adjustment. The next section further discusses the impact of scaling on correction and overthinking.

Parameter Scaling Benefits Knowledge More Than Reasoning

Although scaling law demonstrates scaling improves LLM capabilities (Kaplan et al. 2020), it is unclear scaling benefits knowledge or reasoning. Therefore, the capability improvement from scaling is investigated in Figure 5, with values representing the relative improvement compared to the smallest model in each series. Capabilities of knowledge and reasoning are determined according to Equation 3-5. For instance, the fast thinking accuracy of *Qwen 32B* and its smallest counterpart *Qwen 1.5B* are 79.5% and 53.9%. The knowledge improvement from scaling becomes 25.6%.

Scaling Makes Reasoning Significantly More Prudent and Modestly More Intelligent

Figure 6 demonstrates the different scaling dynamics of overthinking and correction. In both (a) *Qwen* and (b) *LLaMA*, the overthinking rate r_o significantly decreases across all domains, while (c) and (d) show that the correction rate r_c only increases modestly in specific domains.

This phenomenon reveals an asymmetry in the effect of parameter scaling on different reasoning behaviors. The substantial reduction in overthinking rate across all domains suggests that scaling primarily enhances the model's prudence — specifically, its ability to avoid introducing noise. In particular, with larger models possessing greater knowledge capacity, the introduction of erroneous knowledge during COT reasoning is reduced, leading to less overthinking.

In contrast, the relatively modest and domain-specific improvement in correction rate implies that the correction capability of LLMs is less responsive to scaling compared to overthinking. This may reflect the fact that correction involves complex multi-step reasoning, which benefits less from parameter growth alone and require targeted training.

Reasoning Occurs in Deeper Layers of LLMs

The following reveals a phenomenon termed "cognitive hierarchy", indicating that knowledge retrieval and reasoning adjustment operate at different hierarchical layers within neural networks. To quantify this, we employ Centered Kernel Alignment (CKA), a measure of neural network similarity (Cortes, Mohri, and Rostamizadeh 2012; Kornblith et al. 2019).

Figure 7 presents the CKA between the same neural network layer under slow and fast thinking, where a lower CKA indicates reduced similarity for the corresponding layer. (a) and (b) exhibit output CKA for *Qwen 7B* and *LLaMA 8B*, while (c) and (d) exhibits attention CKA.

All CKA curves exhibit an initial plateau in lower layers, followed by a decline in higher layers, indicating that lower layers remain similar across thinking modes, while higher layers diverge. This pattern suggests that both modes share knowledge retrieval in lower layers, but slow thinking additionally engages reasoning adjustment in higher layers. Thus, our results suggest knowledge retrieval and reasoning adjustment are primarily localized to lower and higher layers, respectively.

These findings align with recent studies on knowledge editing (Zhang, Li, and Wu 2024; Meng et al. 2022a,b), which identify knowledge as being primarily located in the lower layers. Our study extends this by demonstrating that reasoning adjustment occurs in the higher layers.

Related Works

CoT To enhance the reasoning capability of LLMs, CoT and its variants, including Tree-of-Thought (ToT), Program-of-Thought (PoT), and Graph-of-Thought, have been proposed (Wei et al. 2022; Yao et al. 2023; Chen et al. 2022; Besta et al. 2024), which endow LLMs with human-like cognitive abilities. Early approaches rely on prompt engineering to explicitly guide CoT generation, but these methods suffer from poor generalizability and heavy dependence on domain expertise. Although techniques like Zero-Shot CoT and Auto-CoT are proposed (Kojima et al. 2022; Zhang et al. 2022), it remains challenging to endow LLMs with the capability to automatically generate high-quality CoT.

Reasoning LLMs The emergence of OpenAI's O1 demonstrates the feasibility of large models autonomously generating high-quality Chain-of-Thoughts (CoTs) (OpenAI 2023). While the implementation details of O1 remain undisclosed, several researchers have conducted extensive investigations into its capabilities (Qin et al. 2024; Huang et al. 2024, 2025b). Initially, researchers propose utilizing external models to supervise and guide the generation of CoT (Cobbe et al. 2021). The most representative methods combine Monte Carlo Tree Search (MCTS) with process-supervised reward models (Zhang et al. 2024; Guan et al. 2025). Although these methods effectively enhance LLMs' performance on complex reasoning tasks, MCTS suffers from prohibitive computational latency (Jiang et al. 2024), and training process-supervised models requires massive CoT datasets with step-by-step annotations, which are extremely costly to obtain (Setlur et al. 2024; Lu et al. 2024; Wang et al. 2023). Subsequent efforts aim to internalize reasoning capabilities within LLMs themselves. Studies find that distilling just 3.9K CoT examples elevates a non-reasoning model's performance to match that of reasoning-specialized models (Min et al. 2024). Recently, DeepSeek shows that outcome-based rewards suffice to enable LLMs to generate high-quality CoT autonomously, significantly reducing the training cost of reasoning-capable LLMs (Guo et al. 2025; Team 2025; Team et al. 2025). Consequently,LLMs possess not only extensive knowledge but also advanced reasoning abilities.

Conclusion

We present a cognition attribution architecture that disentangles knowledge and reasoning in LLMs. In this framework, the cognition of LLMs is decomposed into two phase according to dual-system cognitive theory: knowledge retrieval (Phase 1) and reasoning adjustment (Phase 2). Subsequently, LLMs are prompted to generate answers under different cognitive modes, including fast thinking and slow thinking. The performance under fast thinking determine the capability of knowledge retrieval, while the performance gain attributed to slow thinking quantifies the capability of reasoning adjustment. Ultimately, reasoning adjustment is decomposed into correction and overthinking.

Using this architecture, our findings reveal: (1) reasoning adjustment is domain-specific, benefiting reasoning-intensive domains but potentially harming knowledge-intensive domains; (2) parameter scaling primarily enhances both knowledge retrieval and reasoning adjustment, with reasoning adjustment more significant. Moreover, parameter scaling make reasoning adjustment significantly more prudent in all domains and modestly more intelligent in specific domains. (3) knowledge retrieval and reasoning adjustment are hierarchically organized, residing in lower and higher network layers, respectively. Our architecture not only offers new a perspective on the cognitive properties of LLMs, but also provides insights into some existing research, such as scaling law, hierarchical knowledge editing, and reasoning limitations of small model.

Limitations First, our evaluation is limited to multiple-choice QA; future work should extend the analysis to open-ended generation. Second, while we study 15 models, ultra-large models (greater than 100B) are not included. Third, although we define and quantify knowledge and reasoning based on the dual-system theory, our methods are still simplified and may not fully capture the underlying mechanisms in LLMs. Moreover, this study primarily demonstrates the application of our framework. Although we have observed some interesting phenomena using the framework, due to space limitations, a more in-depth exploration is not feasible at this stage, which remains a direction for future work.

Acknowledgements

We gratefully acknowledge the support provided by Non-communicable Chronic Diseases-National Science and Technology Major Project (2024ZD0522702).

References

Abdin, M.; Aneja, J.; Behl, H.; Bubeck, S.; Eldan, R.; Gunasekar, S.; Harrison, M.; Hewett, R. J.; Javaheripi, M.; Kauffmann, P.; et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

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*.

Amini, A.; Gabriel, S.; Lin, P.; Koncel-Kedziorski, R.; Choi, Y.; and Hajishirzi, H. 2019. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. *arXiv preprint arXiv:1905.13319*.

Besta, M.; Blach, N.; Kubicek, A.; Gerstenberger, R.; Podstawski, M.; Gianinazzi, L.; Gajda, J.; Lehmann, T.; Niewiadomski, H.; Nyczyk, P.; et al. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 17682–17690.

Chen, W.; Ma, X.; Wang, X.; and Cohen, W. W. 2022. Program of thoughts prompting: Disentangling computa-

tion from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Cortes, C.; Mohri, M.; and Rostamizadeh, A. 2012. Algorithms for learning kernels based on centered alignment. *The Journal of Machine Learning Research*, 13(1): 795–828.

Cuadron, A.; Li, D.; Ma, W.; Wang, X.; Wang, Y.; Zhuang, S.; Liu, S.; Schroeder, L. G.; Xia, T.; Mao, H.; et al. 2025. The Danger of Overthinking: Examining the Reasoning-Action Dilemma in Agentic Tasks. *arXiv preprint arXiv:2502.08235*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

Gan, Z.; Liao, Y.; and Liu, Y. 2025. Rethinking External Slow-Thinking: From Snowball Errors to Probability of Correct Reasoning. *arXiv preprint arXiv:2501.15602*.

GLM, T.; Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Zhang, D.; Rojas, D.; Feng, G.; Zhao, H.; et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Guan, X.; Zhang, L. L.; Liu, Y.; Shang, N.; Sun, Y.; Zhu, Y.; Yang, F.; and Yang, M. 2025. rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking. *arXiv preprint arXiv:2501.04519*.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Huang, K.; Guo, J.; Li, Z.; Ji, X.; Ge, J.; Li, W.; Guo, Y.; Cai, T.; Yuan, H.; Wang, R.; et al. 2025a. MATH-Perturb: Benchmarking LLMs' Math Reasoning Abilities against Hard Perturbations. *arXiv preprint arXiv:2502.06453*.

Huang, Z.; Geng, G.; Hua, S.; Huang, Z.; Zou, H.; Zhang, S.; Liu, P.; and Zhang, X. 2025b. O1 Replication Journey–Part 3: Inference-time Scaling for Medical Reasoning. *arXiv preprint arXiv:2501.06458*.

Huang, Z.; Zou, H.; Li, X.; Liu, Y.; Zheng, Y.; Chern, E.; Xia, S.; Qin, Y.; Yuan, W.; and Liu, P. 2024. O1 Replication Journey–Part 2: Surpassing O1-preview through Simple Distillation, Big Progress or Bitter Lesson? *arXiv preprint arXiv:2411.16489*.

Jiang, J.; Chen, Z.; Min, Y.; Chen, J.; Cheng, X.; Wang, J.; Tang, Y.; Sun, H.; Deng, J.; Zhao, W. X.; et al. 2024. Technical report: Enhancing llm reasoning with reward-guided tree search. *arXiv preprint arXiv:2411.11694*.

Kahneman, D. 2011. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux. ISBN 978-0374275631.

Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; and Amodei, D. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Ke, Z.; Shao, Y.; Lin, H.; Konishi, T.; Kim, G.; and Liu, B. 2023. Continual pre-training of language models. *arXiv preprint arXiv:2302.03241*.

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.

Kornblith, S.; Norouzi, M.; Lee, H.; and Hinton, G. 2019. Similarity of neural network representations revisited. In *International conference on machine learning*, 3519–3529. PMLR.

Li, X. L.; and Liang, P. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Li, Y.; Yue, X.; Xu, Z.; Jiang, F.; Niu, L.; Lin, B. Y.; Ramasubramanian, B.; and Poovendran, R. 2025. Small models struggle to learn from strong reasoners. *arXiv preprint arXiv:2502.12143*.

Liu, R.; Geng, J.; Wu, A. J.; Sucholutsky, I.; Lombrozo, T.; and Griffiths, T. L. 2024. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. *arXiv preprint arXiv:2410.21333*.

Liu, X.; Ji, K.; Fu, Y.; Tam, W. L.; Du, Z.; Yang, Z.; and Tang, J. 2021. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*.

Lu, J.; Dou, Z.; Wang, H.; Cao, Z.; Dai, J.; Feng, Y.; and Guo, Z. 2024. Autopsv: Automated process-supervised verifier. *Advances in Neural Information Processing Systems*, 37: 79935–79962.

Meng, K.; Bau, D.; Andonian, A.; and Belinkov, Y. 2022a. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35: 17359–17372.

Meng, K.; Sharma, A. S.; Andonian, A.; Belinkov, Y.; and Bau, D. 2022b. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*.

Min, Y.; Chen, Z.; Jiang, J.; Chen, J.; Deng, J.; Hu, Y.; Tang, Y.; Wang, J.; Cheng, X.; Song, H.; et al. 2024. Imitate, explore, and self-improve: A reproduction report on slow-thinking reasoning systems. *arXiv preprint arXiv:2412.09413*.

Mirzadeh, I.; Alizadeh, K.; Shahrokhi, H.; Tuzel, O.; Bengio, S.; and Farajtabar, M. 2024. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*.

OpenAI. 2023. Learning to Reason with Large Language Models. OpenAI Blog. Accessed: 2024-07-20.

Qin, Y.; Li, X.; Zou, H.; Liu, Y.; Xia, S.; Huang, Z.; Ye, Y.; Yuan, W.; Liu, H.; Li, Y.; et al. 2024. O1 Replication Journey: A Strategic Progress Report–Part 1. *arXiv preprint arXiv:2410.18982*.

Setlur, A.; Nagpal, C.; Fisch, A.; Geng, X.; Eisenstein, J.; Agarwal, R.; Agarwal, A.; Berant, J.; and Kumar, A. 2024. Rewarding progress: Scaling automated process verifiers for llm reasoning. *arXiv preprint arXiv:2410.08146*.

Team, G.; Riviere, M.; Pathak, S.; Sessa, P. G.; Hardin, C.; Bhupatiraju, S.; Hussenot, L.; Mesnard, T.; Shahriari, B.; Ramé, A.; et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Team, K.; Du, A.; Gao, B.; Xing, B.; Jiang, C.; Chen, C.; Li, C.; Xiao, C.; Du, C.; Liao, C.; et al. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

Team, Q. 2025. QwQ-32B: Embracing the Power of Reinforcement Learning. Qwen Blog. Accessed: 2024-07-20.

Ton, J.-F.; Taufiq, M. F.; and Liu, Y. 2024. Understanding Chain-of-Thought in LLMs through Information Theory. *arXiv preprint arXiv:2411.11984*.

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*.

Wang, P.; Li, L.; Shao, Z.; Xu, R.; Dai, D.; Li, Y.; Chen, D.; Wu, Y.; and Sui, Z. 2023. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. *arXiv preprint arXiv:2312.08935*.

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.

Xie, Y.; Aggarwal, K.; and Ahmad, A. 2024. Efficient continual pre-training for building domain specific large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, 10184–10201.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Yang, S.; Zhao, H.; Zhu, S.; Zhou, G.; Xu, H.; Jia, Y.; and Zan, H. 2024b. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 19368–19376.

Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T.; Cao, Y.; and Narasimhan, K. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36: 11809–11822.

Zhang, D.; Zhoubian, S.; Hu, Z.; Yue, Y.; Dong, Y.; and Tang, J. 2024. Rest-mcts*: Llm self-training via process reward guided tree search. *Advances in Neural Information Processing Systems*, 37: 64735–64772.

Zhang, X.; Li, M.; and Wu, J. 2024. Co-occurrence is not factual association in language models. *arXiv preprint arXiv:2409.14057*.

Zhang, X.; Wu, J.; He, Z.; Liu, X.; and Su, Y. 2018. Medical exam question answering with large-scale reading comprehension. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Zhang, Z.; Zhang, A.; Li, M.; and Smola, A. 2022. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.