Skill Decomposition and Composition: A Human-Like Evaluation Framework for Assessing LLMs' Reasoning Abilities

Anonymous ACL submission

Abstract

Large language models (LLMs) have demonstrated remarkable reasoning capabilities across tasks such as commonsense reasoning, mathematical problem-solving, and logical deduction. However, existing evaluation methods, which rely on average accuracy or structured reasoning tasks, provide limited insights into the underlying reasoning mechanisms of LLMs. Correct answers do not necessarily indicate robust reasoning and coarse-grained metrics fail to guide meaningful improvements in reasoning performance. To address this, we propose a human-like reasoning evaluation framework inspired by skill decomposition and skill composition—key cognitive processes in human problem-solving. Specifically, we first annotate the required skills using LLMs and then employ these skills to evaluate the fine-grained reasoning capabilities of LLMs. Our framework refines evaluation metrics by transitioning from accuracy-based measures to skill-level assessments, providing deeper insights into LLMs' reasoning processes from a human-like perspective. Experiments on diverse benchmarks reveal critical insights into LLMs' reasoning strengths and limitations, highlighting the importance of granular evaluation. Code is available at `https://anonymous.4open.science/r/SkillDeCo-76ED/`.

1 Introduction

Large language models (LLMs) (Achiam et al., 2023; Team et al., 2024; Touvron et al., 2023; DeepSeek-AI, 2025) have demonstrated outstanding capabilities in reasoning across commonsense task, mathematical problem, and logical reasoning (Yu et al., 2024; Lai et al., 2024; Huang and Chang, 2022). Recent advancements in reasoning have focused on innovative prompting strategies. Specifically, studies such as Chain-of-Thought (CoT) (Wei et al., 2022) and Tree-of-Thought (ToT) (Yao et al., 2023) have pioneered methods to improve LLMs' reasoning by structuring intermediate reasoning steps. OpenAI's o1 (OpenAI, 2024b) series models further advance reasoning by introducing inference-time scaling, extending the length of CoT reasoning processes to achieve more powerful and nuanced reasoning behaviors. Additionally, DeepSeek-R1 (DeepSeek-AI, 2025) employing GRPO (Shao et al., 2024a) as the reinforcement learning framework achieves competitive performance across math and code reasoning tasks. The human-like intelligence demonstrated by LLMs in reasoning tasks has sparked significant interest in comprehensively evaluating their reasoning ability. Existing studies evaluate the reasoning ability of LLMs by testing the correctness of their responses to test samples or the outputs of structured reasoning tasks (Talmor et al., 2019; Liu et al., 2023; Cobbe et al., 2021; Chollet, 2019). However, these evaluations are based on the coarse-grained metric of average accuracy of LLM responses, and *correct doesn't mean the LLM can reason*. Therefore, they fail to meet the need for a deep understanding of the underlying reasoning mechanisms of LLM reasoning.

To gain deeper insights into the human-like intelligence exhibited by LLMs in problem-solving, it is crucial to explore whether their reasoning mechanisms resemble the psychological processes of human problem-solving. Psychological research has shown that skill decomposition and skill composition are fundamental reasoning abilities for humans to tackle complex problems (Müller and Sternad, 2004; Frederiksen and White, 1989; Smalley et al., 2001). As illustrated in Figure 1(a), when answering questions, the human brain subconsciously decomposes the problem and maps it to several acquired skills, such as “time conversion" and “hourly rates". Based on these decomposed skills, humans implicitly combine past experiences to solve complex problems, as shown in Figure 1(b). This inspires us to explore whether LLMs possess similar abilities in skill decomposition and composition to

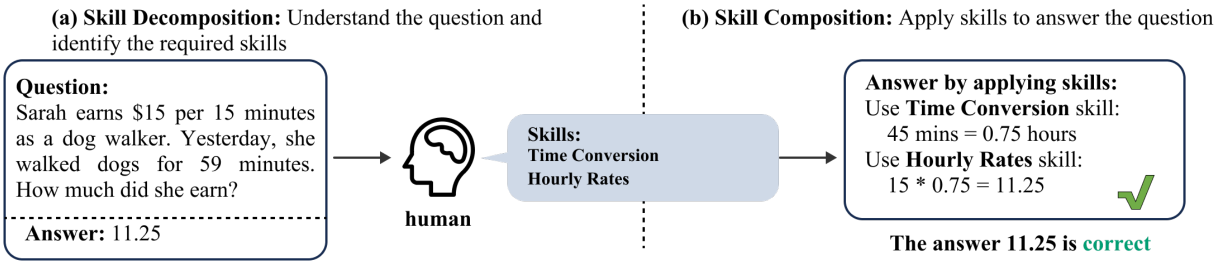

Figure 1: An example of the human reasoning process, comprising (a) **Skill Decomposition** and (b) **Skill Composition**. The whole process begins with skill decomposition, where humans address a problem by first comprehending the question and identifying the required skills (e.g., knowledge concepts such as *Time Conversion* and *Hourly Rates*) for its solution. Then, these skills are composed to arrive at the correct answer, as depicted in (b).

address complex tasks.

Based on these considerations, we propose a human-like reasoning evaluation framework aimed at assessing LLMs' skill decomposition capabilities and their reasoning abilities through skill composition. This is an open-ended problem, and its key challenge lies in the black-box nature of LLM reasoning, making it difficult to observe their internal processes. To address this, we first design a human-like thinking pipeline that guides LLMs to explicitly perform skill decomposition and composition. During the skill decomposition phase, we use structured prompts to guide LLMs to analyze problems and identify required skills based on their pre-trained knowledge. Subsequently, in the skill composition phase, LLMs invoke and process their pre-trained knowledge to solve problems.

Along this pipeline, we evaluate the human-like reasoning abilities of LLMs by testing their performance in task skill decomposition and skill composition to answer questions using standard datasets. Evaluating skill decomposition remains challenging due to the varying ways LLMs understand and process tasks, leading to differences in the granularity of their skill decompositions. For example, as shown in Figure 1, a problem may be decomposed into “*time conversion*” and “*hourly rates*”, which can be further broken down into finer-grained skills such as *multiplication*”, *fractions*”, and “*logic*”. As a result, evaluating skill decomposition using a fixed standard is difficult. To address this issue, we introduce the concept of “skill annotation”. The core idea is to identify the required skills for each evaluation instance and map both the skills and the ground truth into a unified semantic space with abstract granularity, ensuring comparability. Specifically, before evaluation, we use an additional LLM to decompose each evaluation instance and collect all possible skills, which are then clustered into higher-level semantic skills, forming an advanced skill pool. During the evaluation, both the skills decomposed by the LLMs and the ground truth are semantically aligned through the skill pool, enabling a more accurate assessment of skill decomposition.

We apply the proposed evaluation framework to conduct extensive experiments across different LLMs, assessing their skill decomposition, reasoning after skill composition, and the impact of human-like reasoning processes on overall performance. Through in-depth analysis of the experimental results, we uncover novel insights from the human-like reasoning evaluation. For instance, some models exhibit similar performance on traditional answer accuracy metrics, yet show a significant disparity in skill decomposition performance. This discrepancy suggests that traditional evaluation metrics may overestimate a model's reasoning capabilities, as correct answers may be memorized during training rather than derived through genuine reasoning. We also demonstrate that by showing some skills the reasoning ability can be improved. Our main contributions are as follows:

- Skill Annotation for Reasoning Dataset: We release the extended datasets with the skill annotation and the code to annotate new benchmarks. This resource enables finer-grained analysis and benchmarking of reasoning capabilities.
- New Evaluation for LLM Reasoning: We present a new human-like framework and metrics for evaluating LLM reasoning ability from skill decomposition, and skill composition perspectives at a fine-grained skill level.
- Experimental Key Findings: Through experiments, we identify significant variations in skill-level performance across reasoning tasks, offering a detailed understanding of the strengths and limitations of current LLMs.

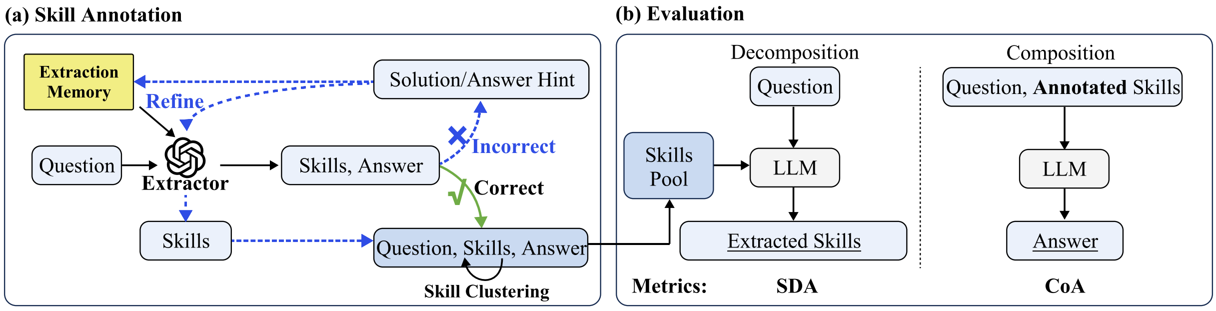

Figure 2: Our human-like reasoning evaluation framework. (a) The skills annotation framework for annotating ground-truth skills for each sentence. The black line represents the initial identification, while the *blue* dashed line indicates re-identification when the initial answer is incorrect. (b) The evaluation process for assessing the ability to decompose and compose skills

2 Human-like Reasoning Evaluation

To gain deeper insights into the human-like intelligence exhibited by LLMs in problem-solving, it is crucial to investigate whether their reasoning mechanisms align with the cognitive processes humans employ to solve problems. In this work, we draw inspiration from the subconscious skill decomposition and composition processes that humans use when answering questions. Specifically, we evaluate LLMs in three key aspects: (1) their ability to decompose tasks into constituent skills, (2) their ability to compose these skills to solve problems, and (3) their overall predictive performance when guided by human-like reasoning processes.

To conduct this evaluation, we propose a pipeline that guides LLMs to think in a human-like manner (see Section 2.2). Within this pipeline, we design three metrics to assess the aforementioned capabilities. Furthermore, due to semantic biases, directly evaluating the skills decomposed by LLMs can be challenging (as mentioned in the introduction). To address this issue, we introduce an additional skill annotation step before evaluation (see Section 2.1). By constructing a semantically abstract skill pool, we map the skills outputted by LLMs and the ground truth into a unified semantic space with the same level of granularity, ensuring comparability in measurement. Note that all the prompts introduced in this section are provided in Appendix A.

2.1 Skill Annotation

The skill annotation step is introduced before the evaluation to address the challenge of comparing the semantic granularity between LLM outputs and ground truth during skill decomposition evaluation. Specifically, we use GPT-4o (OpenAI, 2024a) as the annotator to pre-collect possible skills. These skills are then clustered to form a semantically abstract skill pool. By constructing this pool, we map both the skills outputted by LLMs and the ground truth into a unified semantic space with the same level of granularity, ensuring comparability in measurement.

2.1.1 Skill Extraction

Modern benchmarks for evaluating LLM reasoning ability typically consist of question-answer pairs: $D = \{(q_1, a_1), (q_2, a_1), ..., (q_n, a_n)\}$. Given an evaluation instance (q_i, a_i), our goal is to construct the skill set S_i required to solve the question q_i and produce the correct answer a_i. In our framework, each skill $s_m \in S_i$ is represented as a tuple $s_m = (\text{name}, \text{description}, \text{usage})$, where name denotes the skill name, description provides a detailed explanation of the skill, and usage offers a representative example of its application.

To ensure the reliability of annotated skills, the skill set S_i is obtained as below.

Initially, the LLM extractor is prompted to identify the skill set S_i required to solve the question, along with the reasoning process using these skills, and the final answer a'_i:

$$(S_i, a'_i) = \text{LLM}_{\text{GPT-4o}}(q_i, p_{\text{ide}}),$$

where p_{ide} represents the prompt to identify the skill set. Only the skill set S_i that yields the correct answer a'_i is retained for the evaluation instance. In cases where a'_i is incorrect, the reasoning process based on S_i may be unreliable. To prevent the LLM extractor from repeating similar mistakes and wasting resources, we employ a self-summarization mechanism (Matelsky et al., 2023) to guide the LLM extractor in reflecting on its errors and summarizing its experiences. These experiences are then stored in the **Extraction Memory** module as annotation knowledge, as illustrated by the *blue* dashed line in Figure 2(a).

Building on the knowledge accumulated in the Extraction Memory, the LLM synthesizes insights from failed extractions to enhance its annotation process. Specifically, the extractor is provided with additional hints, such as the correct solution or answer, to reanalyze the problem and generate a more reliable skill set, S_i^{Re}:

$$(S_i, a_i') = \text{LLM}_{\text{GPT-4o}}(q_i, a_i, M_{fail}, p_{\text{Re}}),$$

where p_{Re} is a reidentification prompt incorporating the original question, solution hints, and failed extraction experiences. This iterative refinement ensures that each evaluation instance is ultimately annotated with a robust and accurate skill set, improving both the interpretability and reliability of the analysis.

2.1.2 Skill clustering

Following the skill annotation process, we obtain a labeled set of skills for each evaluation instance. However, the sheer volume of skills poses a challenge due to the presence of semantically equivalent or highly similar skills (e.g., "*Basic Arithmetic*" and "*Basic Arithmetic Operation*"). To address this issue, we introduce a systematic skill clustering approach designed to reduce redundancy and create a semantically coherent skill taxonomy. Once the skill taxonomy is established, each skill in the skill set S_i is mapped to its corresponding clustered skill category. The skill clustering process is implemented through a two-step approach: Batch Clustering and Post-Processing. This methodology balances computational efficiency with semantic accuracy, leveraging the strengths of LLMs to handle both large-scale data and nuanced semantic relationships.

Batch Clustering In the first step, the skill set is partitioned into smaller, manageable batches to enable parallelized processing. Within each batch, an LLM categorizes the skills based on their descriptive text, generating the following outputs for each cluster:{*Category Names, Skill Lists, Brief Descriptions, Representative Usages*}. This step addresses the computational limitations of LLMs when processing long texts, ensuring efficient handling of large skill sets.

Post-Processing The second step refines the initial clusters by assessing and merging semantically similar categories. An LLM evaluates the category descriptions and representative usages generated during the Batch Clustering phase, identifying and merging clusters that exhibit conceptual alignment. This post-clustering adjustment ensures that the final skill taxonomy is both logically consistent and semantically meaningful.

This two-step clustering process not only addresses the challenge of skill redundancy but also provides a robust framework for organizing and interpreting large-scale skill datasets. The resulting taxonomy serves as a foundation for further analysis, enabling a more human-like assessment of LLMs' reasoning abilities by aligning skill representations with cognitively plausible structures.

2.2 Skill-based Evaluation

Building upon the reasoning benchmark established through skill annotation, we then delve into understanding language model reasoning by evaluating the skill decomposition and composition abilities of LLMs.

Skill Decomposition Ability Skill decomposition refers to the ability to analyze a problem and identify the specific skills, knowledge, or sub-tasks required to solve it.

In our framework, we prompt the LLM to emulate human-like reasoning by identifying the skills required to solve a given problem. For each problem, the LLM retrieves the relevant skills from a predefined *skill pool*—a comprehensive set of skill categories derived through a systematic skill clustering process (See Section 2.1.2). Formally, this retrieval process is represented as:

$$S_i^{de} = \text{LLM}_{\text{eval}}(q_i, p_{\text{ide}}),$$

where S_i^{de} is the skill set required for question q_i, $\mathcal{S}$ represents the global skill pool obtained after clustering, p_{ide} is the prompt to identify the required skills. and LLM_{eval} is the LLM to be evaluated. The skill identification process effectively simulates how humans approach reasoning tasks by breaking them into constituent skills. Then, we use $S_i^{map} = \text{LLM}_{\text{eval}}(S_i^{de}, p_{\text{retrieve}})$, to map the identified skills to skill clusters, thereby avoiding semantic discrepancies and enabling quantitative assessment. Here, p_{retrieve} is the prompt used to guide the LLM in retrieving the most relevant skills from the skill pool. To evaluate the effectiveness of the skill decomposition process, we introduce the **S**kill **D**ecomposition **A**ccuracy (***SDA***) metric. The

SDA metric quantifies how well the LLM-retrieved skill set S_i^{map} matches the ground-truth skill set S_i^{true}, which is established through expert skill annotation. The metric is formally defined as:

$$SDA = \frac{\sum_{i=1}^{n} |S_i^{map} \cap S_i^{\text{true}}|}{\sum_{i=1}^{n} |S_i^{\text{true}}|}.$$

Skill Composition Ability Skill composition refers to the capability to systematically integrate identified skills or sub-tasks to derive a coherent and correct solution. This ability is fundamental for solving complex problems that require a structured and sequential reasoning process.

In our framework, the skill composition ability of the LLM is evaluated by presenting it with a problem q_i alongside its ground-truth skill set S_i^{true}. The LLM is prompted to utilize these skills step by step to construct a solution a_i^{Com}. Formally, this process is expressed as:

$$a_i^{\text{Com}} = \text{LLM}_{\text{eval}}(q_i, S_i^{\text{true}}, p_{\text{compose}}),$$

where p_{compose} is a structured prompt designed to guide the LLM in leveraging the provided skill set for systematic problem-solving.

To evaluate the effectiveness of the LLM's skill **Co**mposition **A**bility, we introduce the skill composition accuracy (***CoA***) metric. The *CoA* quantifies the alignment between the generated solution a_i^{Com} and the ground-truth solution a_i, and is formally defined as: $CoA = \frac{\sum_{i=1}^{n} \mathbb{I}(a_i^{\text{Com}} = a_i)}{n}$.

To conclude, the *SDA* metric offers an objective measure of the alignment between the LLM's skill decomposition and the annotated ground-truth skill sets. Similarly, the *CoA* metric provides a quantitative assessment of the LLM's ability to effectively compose and apply the identified skills to derive accurate solutions. It is important to note that we also decompose and then compose the reasoning process to derive the final answer, enabling an investigation of the entire human-like reasoning process. The final answer is used to calculate the **D**ecompose-and-**C**ompose **A**ccuracy (*DCA*), which serves as a measure of reasoning ability. The DCA metric reveals the impact of human-like reasoning processes on overall performance. Together, these three metrics offer a rigorous framework for evaluating the human-like reasoning capabilities of LLMs.

3 Experiment

In this section, we aim to answer three research questions (RQs):

- **RQ1:** How does human-like reasoning evaluation offer insights beyond traditional *ACC* metrics?
- **RQ2:** How do LLMs perform across different reasoning skills?
- **RQ3:** Can skill-specific information enhance the reasoning performance of LLMs?

3.1 Experimental Setup

Utilized LLMs Taking into account both the reasoning capabilities and the cost-effectiveness of the model, we select GPT-4o-08-06 (OpenAI, 2024a) to annotate skills as the ground-truth skills for existing benchmarks. Furthermore, we evaluate the skill decomposition and composition abilities of five mainstream LLMs, including three closed-source models: GPT-4 (Achiam et al., 2023), GPT-3.5-Turbo (OpenAI, 2023) and Gemini 1.5 pro (Team et al., 2024); as well as two open-source models: Llama3-8B (Touvron et al., 2023) and Qwen2-7B-Instruct (Shao et al., 2024b).

Datasets Our skill annotation and human-like reasoning evaluation is conducted on three datasets: (1) a language understanding and reasoning dataset, **LogiQA 2.0** (Liu et al., 2023); (2) mathematical reasoning datasets, **MATH** (Hendrycks et al., 2021) and **GSM8K** (Cobbe et al., 2021). The LogiQA 2.0 test set consists of 1,572 question-answer pairs covering 5 types of complex logical reasoning. The MATH test set contains 5,000 mathematical problems across 5 difficulty levels and 7 subjects. The GSM8K test set includes 1,319 mathematical problems, each accompanied by a step-by-step solution and a ground-truth answer.

Evaluation Metrics We aim to derive insights from a human-like evaluation framework by contrasting it with traditional evaluations based solely on final answer accuracy (*ACC*). Specifically, for skill decomposition, we report the average skill decomposition accuracy (*SDA*) to quantify how well the LLM-retrieved skill set matches the ground-truth skill set, thereby demonstrating the effectiveness of the skill decomposition ability. We also report *DCA* to show the entire human-like Decompose-and-Compose reasoning performance. For composition, we report the skill composition accuracy metric (*CoA*) to quantify the alignment between the generated solution based on the ground-truth skills and the ground-truth solution.

Table 1: Evaluation results on different LLMs.

Models	LogiQA 2.0				Math				GSM8K			
	ACC↑	SDA ↑	DCA ↑	CoA↑	ACC↑	SDA↑	DCA ↑	CoA ↑	ACC↑	SDA↑	DCA↑	CoA↑
GPT-4-1106	69.7	32.3	67.6	72.3	60.2	51.0	57.6	59.5	89.2	70.6	88.8	89.4
GPT-3.5-turbo	54.3	17.3	52.4	58.3	40.2	42.7	44.3	48.6	79.3	52.2	66.0	84.3
Gemini-1.5 pro	75.4	30.7	74.0	77.7	81.0	55.2	81.2	82.0	90.8	66.5	93.0	94.5
Llama-3.1-8B-Instruct	53.8	32.9	52.4	60.1	47.9	45.9	39.6	47.0	84.5	39.4	57.9	82.2
Qwen2-7B-Instruct	53.4	16.2	52.7	56.1	48.6	44.1	43.5	50.7	62.1	63.6	67.2	70.3

3.2 Main Results (RQ1)

After annotating skills for each reasoning evaluation benchmark, we conduct a human-like reasoning evaluation method introduced in Section 2.2 on LogiQA 2.0, MATH, and GSM8K datasets. From Table 1's results, we have the following interesting and valuable findings:

• **Finding 1:** The results for the two metrics assessing skill decomposition ability, *SDA* and *DCA*, show a strong correlation across different LLMs. This indicates that if a model can accurately extract the correct set of skills required for a problem, its answer accuracy also tends to be higher. However, when comparing traditional metrics like *ACC* to *SDA*, discrepancies arise. For example, on the LogiQA 2.0 dataset, the *ACC* values of Qwen2-7B-Instruct and Llama-3.1-8B-Instruct are close, while their *SDA* values differ by 16%. This disparity suggests that the models may not effectively analyze the skills required to solve the problem. Actually, inferior *SDA* but high *ACC* might stem from memorizing the correct answers during training rather than employing true reasoning improvements. These observations underscore the importance of evaluating LLMs' decomposition ability and the novel perspective introduced by *SDA* in reasoning evaluation. Traditional answer accuracy metrics like *ACC* may overestimate a model's reasoning capabilities, where some correct answer is driven by memorization or dataset-specific biases. The findings further highlight that enhancing skill decomposition ability should be a potential focus in model training and fine-tuning, as it directly impacts both reasoning accuracy and generalization.

• **Finding 2:** The second experimental finding reveals that *CoA*, which measures the alignment between model-generated solutions informed by annotated skill labels and the ground-truth label accuracy, consistently outperforms both *ACC* and DCA. This demonstrates the reliability of the ground-truth skills generated by our automatic skill annotation method, as reflected in improved performance. Furthermore, this observation highlights the crucial role that skill labels play in enhancing LLM reasoning abilities. For instance, the Llama-3.1-8B-Instruct model, which performs worse than GPT-3.5-turbo in *ACC*, outperforms it in *CoA*. This suggests that the inclusion of well-defined skill labels can improve a model's ability to generate accurate solutions, even when it struggles with decomposing complex tasks.

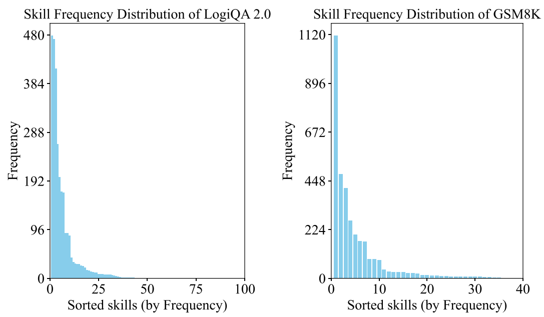

Figure 3: Frequency distribution of the extracted skills.

• **Finding 3:** Upon analyzing the logical reasoning performance on LogiQA 2.0, it is evident that all models struggle with the decomposition task, showing low *SDA* scores ranging from 17% to 33%. In contrast, their performance in mathematical reasoning on the MATH dataset is higher, ranging from 40% to 50%. However, the traditional *ACC* scores are similar across both benchmarks. This performance disparity can likely be attributed to the more structured and rule-based nature of mathematical problems, which aligns better with the models' capabilities for decomposition and skill alignment. In contrast, logical reasoning tasks demand abstract thinking and complex inferences, which are inherently harder to decompose due to their implicit and multifaceted nature. These findings underscore the need for further advancements in LLMs' reasoning capabilities, especially in abstract and non-quantitative domains, to improve their overall performance across a range of tasks.

3.3 Skill-level Reasoning Analysis (RQ2)

3.3.1 Skills Skewed Distribution

We have plotted the distribution of the extracted skills in Figure 3, showing the sorted usage frequency of skills in the LogiQA 2.0 and GSM8K benchmarks. The total number of instances is similar but the total number of skills required is quite different, indicating that LogiQA 2.0 demands a greater variety of specific skills. Furthermore, the skill distribution is highly skewed, especially in datasets like GSM8K, where 10 skills account over 90% of the test cases. This concentration of skills in certain areas leads to an uneven distribution of skill usage across different problems. To explore this further, we report the average *SDA* scores for the top 10 and bottom 10 most frequently used skills in GPT-4-11-06 is 72.3% and 6% respectively. Notably, the model performs worse on the bottom 10 skills, indicating that its ability to handle less frequent but equally important skills is limited. This suggests that GPT-4 may struggle with edge cases or rare skills that are critical for certain problems. These findings emphasize the need for LLMs to adopt a more balanced training approach, one that ensures the proper handling of rare or underrepresented skills.

3.3.2 Skill-level Reasoning Performance

Skill annotation plays a crucial role in advancing fine-grained reasoning assessment. In this study, we show 5 skills including {*Algebra, Fractions, Comparison and Analysis, Cost Calculation, Financial Calculation*} from the GSM8K dataset, and report the decomposition performance using *SDA* composition performance using *CoA* in Figure 4 across three closed-source models: GPT-3.5 Turbo and Gemini-1.5 Pro.

For the five skills, we observe that different models excel at distinct tasks in both decomposition and composition. Notably, Gemini-1.5-pro outperforms GPT-3.5-turbo across the board in composition tasks, demonstrating superior performance. However, in decomposition, Gemini-1.5-pro lags behind in *Financial Calculation*. This suggests that Gemini places greater emphasis on enhancing the model's ability to identify the required skills for specific scenario-based questions. In composition, GPT-3.5-turbo struggles with applying the *Cost Calculation* skill to solve the question, indicating that it should focus more on improving its application of this skill.

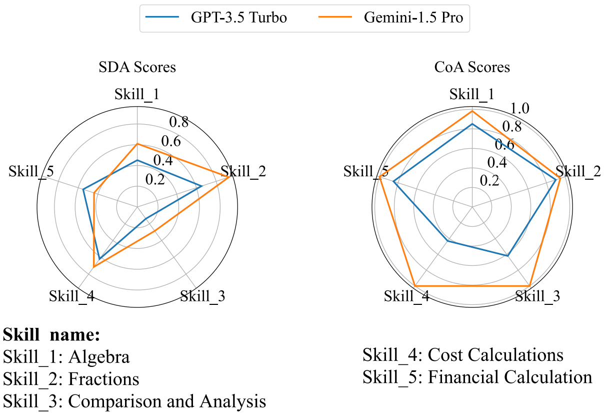

Figure 4: Performance Comparison of GPT-3.5-Turbo and Gemini-1.5 Pro on Decomposition (SDA) and Composition (CoA) for different Skills on GSM8K.

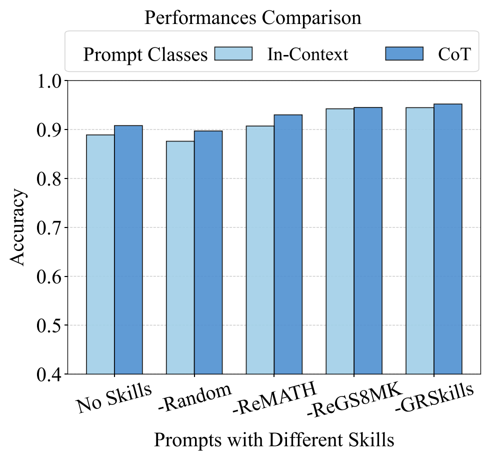

Figure 5: Performances comparison of direct and CoT prompts with Different Skills.

3.4 Skills Help LLM Reasoning (RQ3)

To investigate the impact of skill-specific information on LLM reasoning, we compare direct prompt In-context-class (our prompting method adopted in the main results) and CoT-class on the GSM8K dataset. Within each class, we provide skill information using four approaches: random skills from GSM8K's pool (-"random"), skills retrieved from the MATH skill pool (-"ReMATH"), skills retrieved from the GSM8K skill pool (-"ReGSM8K"), and the instance's ground-truth skills (-"GRSkills"). The results, as shown in Figure 5, illustrate how different sources of skill information influence reasoning performance.

First, we observe that across different skill hints, "-GRSkills" consistently achieves the best performance. Interestingly, in-context prompting with ground-truth skills performs competitively with

CoT, indicating that providing reliable skill information can effectively guide LLMs to adopt step-by-step reasoning and enhance their reasoning capabilities. Second, we note that "-ReMATH" outperforms both the no-skill baseline and the -random" setting, demonstrating that the MATH-extracted skill pool can be leveraged to improve performance on GSM8K. This finding highlights the transferability of skill-based knowledge across different benchmarks, emphasizing the potential of cross-task skill utilization for enhancing LLM reasoning.

4 Related Works

4.1 LLMs Reasoning

Reasoning is a core facet of intelligence, critical for tasks such as decision-making and solving complex problems like mathematics (Yu et al., 2024). Recent advancements have centered around innovative prompting strategies to enhance LLM reasoning capabilities. Methods such as Chain-of-Thought (CoT) (Wei et al., 2022) and Tree-of-Thought (ToT) (Yao et al., 2023) have advanced LLMs by structuring intermediate reasoning steps. OpenAI's O1 series (OpenAI, 2024b) further refines advance reasoning with inference-time scaling, extending CoT reasoning to generate more nuanced outcomes. Moreover, DeepSeek-R1 (DeepSeek-AI, 2025), using the GRPO framework (Shao et al., 2024a), achieves performance comparable to OpenAI-O1 on reasoning tasks. Existing studies (Huang et al., 2024; Wei et al., 2022; DeepSeek-AI, 2025; OpenAI, 2024b) primarily assess reasoning ability by using benchmarks such as commonsense reasoning (Talmor et al., 2019), math reasoning (Hendrycks et al., 2021; Fan et al., 2024) and Strategic Reasoning (Zhang et al., 2024), where questions are presented as input, and the models' performance is evaluated based on the average accuracy of their responses. This coarse evaluation approach provides minimal insights into the detailed analysis of LLMs' reasoning ability and offers limited guidance for improving their reasoning performance. To address this issue, this work focuses on analyzing mathematical reasoning at the skill level from a human-like perspective (Johnson and Kuennen, 2006).

4.2 Skill in Reasoning

Skills are foundational to educational assessment, serving as measurable indicators of a learner's ability to apply knowledge, judgment, and techniques within specific domains (Smee, 2003). The importance of skills in enhancing the reasoning abilities of LLMs has gained attention through approaches like prompting (Didolkar et al., 2024) and fine-tuning with skill-focused data synthesis (Chen et al., 2024b; Huang et al., 2024). For instance, Didolkar et al. employs prompts to guide LLMs toward identifying and applying individual skills, improving mathematical reasoning. However, a comprehensive framework for assessing LLM reasoning at the skill level remains underdeveloped. Recent efforts by Chen et al. leverage GPT-4 to generate rationales that represent underlying skills in general benchmarks, marking a significant step forward. This work addresses the critical gap by integrating skill-level analysis aligned with the human-like reasoning process, enabling more granular reasoning assessments and offering targeted instructional insights to enhance requisite skills.

5 Conclusion

In this paper, we introduced a novel human-like reasoning evaluation framework designed to assess the reasoning capabilities of large language models (LLMs) by focusing on skill decomposition and skill composition, which closely align with human cognitive processes. To be specific, unlike traditional evaluation methods that depend on average accuracy metrics, our framework offers deeper insights into the underlying reasoning mechanisms of LLMs. To achieve this, we first explicitly annotated the skills required to solve a given problem. During the evaluation, we assessed LLMs' reasoning abilities through two key dimensions: skill decomposition (identifying the necessary skills to address a problem) and skill composition (integrating these skills to generate coherent solutions). The experimental results revealed new insights from the human-like reasoning evaluation. Specifically, while some models performed similarly on traditional answer accuracy metrics, they showed significant differences in skill decomposition performance. This discrepancy suggests that traditional evaluation metrics may overestimate a model's reasoning capabilities, as correct answers could be memorized during training rather than generated through reasoning. Besides, our findings demonstrate that incorporating skills can enhance the reasoning abilities of LLMs.

6 Limitations

In this paper, we propose a human-like reasoning evaluation framework by emphasizing skill decomposition and skill composition based on the annotated skills. However, in practice, skills commonly exhibit multi-level and complex structural relationships when solving complex problems. The method proposed in this paper is an initial exploration of the basic framework for human-like reasoning evaluation and does not delve deeply into the hierarchical relationships and interactions among skills. For example, certain skills (e.g., *Functions*) may depend on other sub-skills (e.g., *Numbers*, *Addition*, and *Arithmetic*). These intricate structural relationships could affect the accuracy and comprehensiveness of the evaluation results. In the future, we will further investigate the hierarchical structure of skills to more precisely simulate human reasoning mechanisms and enhance the robustness of the evaluation framework.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Feng Chen, Chenhui Gou, Jing Liu, Yang Yang, Zhaoyang Li, Jiyuan Zhang, Zhenbang Sun, Bohan Zhuang, and Qi Wu. 2024a. Evaluating and advancing multimodal large language models in ability lens. *arXiv preprint arXiv:2411.14725*.

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

François Chollet. 2019. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Aniket Didolkar, Anirudh Goyal, Nan Rosemary Ke, Siyuan Guo, Michal Valko, Timothy Lillicrap, Danilo Rezende, Yoshua Bengio, Michael Mozer, and Sanjeev Arora. 2024. Metacognitive capabilities of llms: An exploration in mathematical problem solving. *arXiv preprint arXiv:2405.12205*.

Jingxuan Fan, Sarah Martinson, Erik Y Wang, Kaylie Hausknecht, Jonah Brenner, Danxian Liu, Nianli Peng, Corey Wang, and Michael P Brenner. 2024. Hardmath: A benchmark dataset for challenging problems in applied mathematics. *arXiv preprint arXiv:2410.09988*.

John R Frederiksen and Barbara Y White. 1989. An approach to training based upon principled task decomposition. *Acta psychologica*, 71(1-3):89–146.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *NeurIPS*.

Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*.

Yiming Huang, Xiao Liu, Yeyun Gong, Zhibin Gou, Yelong Shen, Nan Duan, and Weizhu Chen. 2024. Key-point-driven data synthesis with its enhancement on mathematical reasoning. *arXiv preprint arXiv:2403.02333*.

Marianne Johnson and Eric Kuennen. 2006. Basic math skills and performance in an introductory statistics course. *Journal of statistics education*, 14(2).

Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangru Peng, and Jiaya Jia. 2024. Step-dpo: Step-wise preference optimization for long-chain reasoning of llms. *arXiv preprint arXiv:2406.18629*.

Hanmeng Liu, Jian Liu, Leyang Cui, Zhiyang Teng, Nan Duan, Ming Zhou, and Yue Zhang. 2023. Logiqa 2.0—an improved dataset for logical reasoning in natural language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Jordan K Matelsky, Felipe Parodi, Tony Liu, Richard D Lange, and Konrad P Kording. 2023. A large language model-assisted education tool to provide feedback on open-ended responses. *arXiv preprint arXiv:2308.02439*.

Hermann Müller and Dagmar Sternad. 2004. Decomposition of variability in the execution of goal-oriented tasks: three components of skill improvement. *Journal of Experimental Psychology: Human Perception and Performance*, 30(1):212.

OpenAI. 2023. Chatgpt.

OpenAI. 2024a. Hello gpt-4. `https://openai.com/index/hello-gpt-4o/`. Accessed: 2025-02-11.

OpenAI. 2024b. Learning to reason with LLMs. `https://openai.com/index/learning-to-reason-with-llms/`. Accessed: [Insert Date].

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024a.

Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. 2024b. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Regina L Smalley, Mary K Ruetten, and Joann Kozyrev. 2001. *Refining composition skills: Rhetoric and grammar*. Heinle & Heinle Boston, MA.

Sydney Smee. 2003. Skill based assessment. *Bmj*, 326(7391):703–706.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. 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)*, pages 4149–4158.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Fei Yu, Hongbo Zhang, Prayag Tiwari, and Benyou Wang. 2024. Natural language reasoning, a survey. *ACM Computing Surveys*, 56(12):1–39.

Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Adrian de Wynter, Yan Xia, Wenshan Wu, Ting Song, Man Lan, and Furu Wei. 2024. Llm as a mastermind: A survey of strategic reasoning with large language models. *arXiv preprint arXiv:2404.01230*.

A Prompts in Human-like Reasoning Evaluation

The prompts used for skill annotation and decomposition/composition evaluation are presented in Figure 6 (a)-(d) respectively.

Identify Prompt

```
Below is a math question for you to solve.
Your goal is to try your best to solve the problem and provide a response.
You must think step by step and give us a response in STRICT accordance
with our guidelines.
### Question: {question text}

### Response:
   Let's think step by step. Carefully analyze the problem.
   Identify the skills that are used to solve it, skills like a math concept, a
theorem by the name. For each identified skill, provide a brief description of
its role in solving the problem. Next, explain how each skill is applied in the
solution process, ensuring that each skill is used correctly and effectively.
   Do remember to give us a response in STRICT accordance with our
guidelines and your response should be STRICTLY formatted as:
   {{
      "Skills": [
         {{
            "skill_name": "<skill_1>",
            "description": "<brief description of the skill and its role>",
            "usage": "<how the skill is applied in the solution process>"
         }},
         ...
      ],
      "Answer": "<your reasoning process and the final answer based on the
skills and
       the problem itself>"
   }}
```

(a) Prompt to identify the skill set from the question text

Reidentify Prompt

```
Below is a math question for you to solve.
Your goal is to try your best to solve the problem and provide a response.
You must think step by step and give us a response in STRICT accordance
with our guidelines.
### Question: {question text}

### Solution: {solution}

### Extraction Memory: {extraction memory}

### Response:
   Let's think step by step. Carefully analyze the problem and its solution.
   Base on extraction memory, identify the skills that are used to solve it,
skills like a math concept, a theorem by the name. For each identified skill,
provide a brief description of its role in solving the problem. Next, explain
how each skill is applied in the solution process, ensuring that each skill is
used correctly and effectively.
   Do remember to give us a response in STRICT accordance with our
guidelines and your response should be STRICTLY formatted as:
   {{
      "Skills": [
         {{
            "skill_name": "<skill_1>",
            "description": "<brief description of the skill and its role>",
            "usage": "<how the skill is applied in the solution process>"
         }},
         ...
      ],
      "Answer": "<your reasoning process and the final answer based on the
skills and
       the problem itself>"
   }}
```

(b) Prompt to reidentify the skill set from the question text, solution and extraction memory

Decomposition Evaluation Prompt

```
Below is a math question for you to solve.
Your goal is to try your best to solve the problem and provide a response.
You must think step by step and give us a response in STRICT accordance with our
guidelines.
### Question: {question text: }

### Response:
   Let's think step by step. Carefully analyze the problem and identify the skills are
used to solve it from the skill pool including {{skill name, description, usage}}
   Do remember select the the skills from the skill pool:\n{skill_pool}.
   Next, explain how each skill is applied in the solution process, ensuring that each
skill is used correctly and effectively.
   Do remember to give us a response in STRICT accordance with our guidelines and
your response should be STRICTLY formatted as:
   {{
      "Identified Skills": [
         {{
            "skill_name": "<skill_1>",
            "usage":"<how the skill is applied in the solution process>"
         }},
         ...
      ],
      "Answer": "<your reasoning process and the final answer based on the skills and
the problem itself>"
   }}
   """
```

(c) Prompt to retrieve the relevant skills from a skill pool and give a solution based on the retrieved skills

Composition Evaluation Prompt

```
Below is a math question for you to solve.
Your goal is to try your best to solve the problem and provide a response.
You must think step by step and give us a response in STRICT accordance with our
guidelines.
### Question: {question text}

### Skills hint:
   {ground_truth_skills}

### Response:
   Let's think step by step. Carefully analyze the problem and the skills hint, and try
your best to provide a response to solve the problem.
   Next, explain how each skill is applied in the solution process, ensuring that each
skill is used correctly and effectively.
   Do remember to give us a response in STRICT accordance with our guidelines and
your response should be STRICTLY formatted as:
   {{
   "Answer": "<your reasoning process and the final answer based on the skills and
   the problem itself>"
   }}
   """
```

(d) Prompt to guide the LLM in leveraging the annotated skill set for systematic problem-solving

Figure 6: The prompts used for skill annotation and decomposition/composition evaluation.