WAKENLLM: Evaluating Reasoning Potential and Stability in LLMs via Fine-Grained Benchmarking

Anonymous ACL submission

Abstract

Large Language Models (LLMs) frequently output the label *Unknown* in reasoning tasks, where two scenarios may appear: (i) an input sample is genuinely *unverifiable*, but the model cannot understand why; and (ii) a *verifiable* problem that the model fails to solve, and thus the model outputs *Unknown*. We refer to these cases collectively as the *Vague Perception* phenomenon. Current evaluations focus on whether such answers are honest, rather than analyzing the limits of LLM reasoning. To address this, we introduce WAKENLLM, a framework that quantifies the portion of *Unknown* output attributable to model incapacity and evaluates whether stimulation can convert them into either correct answers (*verifiable samples*) or justified explanations (*unverifiable samples*) supported by valid reasoning. Our method offers a clearer picture of the limits of LLM reasoning and the potential for corrections across various datasets. Comprehensive experiments on six LLMs suggest that, without further training or parameter revision, LLMs can achieve up to a 68.53% increase in accuracy on *Vague Perception* samples through guided understanding. Our work reveals that current baseline methods only activate a small portion of LLMs' reasoning potential, indicating considerable unexplored capacity. This extends the empirical upper bounds of reasoning accuracy in LLMs. Consequently, this study deepens our understanding of the latent reasoning capacity of LLMs and offers a new perspective on addressing the *Vague Perception* phenomenon.

1 Introduction

Large Language Models (LLMs) such as GPT-3, PaLM, and GPT-4 have shown impressive zero-shot and few-shot reasoning on benchmarks like BIG-Bench, MMLU, and GSM8K (Brown et al., 2020; OpenAI, 2023; Chowdhery et al., 2022; Srivastava and et al., 2022; Hendrycks and et al., 2021; Cobbe et al., 2021). Their performance is further improved by prompting techniques such as Chain-of-Thought, Self-Consistency, and Program-Aided Reasoning (Kojima et al., 2022; Wei et al., 2022b; Wang et al., 2023c; Lewkowycz and et al., 2022). Meanwhile, most evaluations rank the overall trustworthiness the first and causal atmosphere second(Jiang et al., 2024; Bubeck, 2023). Recent diagnostic benchmarks—such as potential-space reasoning (Hagendorff and Fabi, 2025), visual hallucination stability (Liu et al., 2025), and metaphysical reasoning (MARS) (Wang and Song, 2024)—demonstrate the need for more fine-grained evaluations.

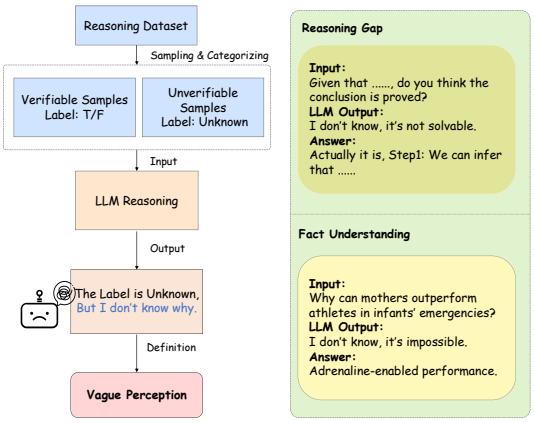

Figure 1: Definition of *Vague Perception*: when an LLM misjudges a verifiable sample as *Unknown*, or it is unable to tell why a sample should be categorized as unknown when facing objectively unsolvable ones. The figure presents two representative reasons attributed to this phenomenon: *Fact Understanding* and *Reasoning Gap*—highlighting 1. Limited ability to solve a problem. 2. Misunderstandings of facts versus failures to link them.

Fine-grained reasoning reliability remains under-explored. Empirical studies have shown that LLMs frequently hallucinate, or output "Unknown" even when the context contains sufficient evidence, while at other times they output *Unknown* without understanding why, facing fundamentally under-specified (Lin, 2023; Zhou et al., 2025) prob-

Table 1: Comparison of different LLM reasoning–guidance frameworks. WAKENLLM integrates all eight dimensions as the first holistic solution for diagnosing and mitigating the *Vague Perception* phenomenon. For used symbols, ✓: Fully covered, ●: Partially covered, ✗: Not covered.

Paradigm	Guidance	Questioning	Stimulation	Dynamic Samples Selection	Reflection	Stage-wise Examining	Fine-Grained Decomposing	Diversity
CoT (Wei et al., 2022b)	✓	✗	✗	✗	✗	✗	✗	✗
ToT (Yao et al., 2023b)	✓	✓	✗	●	●	✓	✓	✓
Faithful-CoT (Lyu et al., 2023)	✓	✓	✓	✗	✗	✓	✗	✓
Plan&Solve (Wang et al., 2023b)	✓	✗	●	✗	✗	✓	✓	✗
Self-Refine (Madaan et al., 2023)	✓	✗	✓	✗	✓	✓	✗	✓
Reflexion (Shinn et al., 2023)	✓	✗	✓	✗	✓	✓	✗	✓
Decomposed Prompting (Khot et al., 2023)	✓	✓	✗	✗	✗	✗	✓	✗
Active Prompt (Diao et al., 2023)	✓	✗	✓	✓	✗	✗	✗	✓
SCORE (Zhang et al., 2024)	✓	✗	✓	✗	✓	✓	✗	✓
ProCo (Wu et al., 2024)	✓	✓	✓	✗	✓	✓	✗	✓
CoTAL (Cohn et al., 2025)	✓	✗	✓	✓	✗	✓	✗	✗
Cochain (Zhao et al., 2025)	✓	✓	✗	✓	✗	✓	✓	✓
ISP^2 (Zhu et al., 2025)	✓	✓	✓	✗	✗	✗	✗	✓
Trace-of-Thought (McDonald and Emami, 2025)	✓	✓	✗	✗	✗	✓	✓	✗
MAPS (de Souza Loureiro et al., 2025)	✓	✗	✓	●	✓	✓	✗	✓
CoD (Xu et al., 2025)	✓	✗	✗	✗	✗	✓	✗	✗
ReSearch (Chen et al., 2025)	✓	✓	✗	✓	✓	✓	✓	✗
WAKENLLM (Ours)	✓	✓	✓	✓	✓	✓	✓	✓

lems. This paper formalizes this phenomenon as *Vague Perception*: a mismatch between a model's reasoning ability and the objective problems. Therefore, a critical gap lies in understanding how LLMs deal with complex reasoning and uncertainty.

In this work, we analyze the root causes of deep-reasoning failures behind *Vague Perception* and propose fine-grained mitigation strategies. To enable systematic measurement, we introduce a new benchmark that spans four subject domains and multiple subcategories, offered in two different formats: (i) *Fact-based benchmark*, where relevant information is listed as discrete atomic statements (e.g., "A is X, B is Y"); and (ii) *Story-based benchmark*, where the same facts are embedded in a coherent passage. This design probes how presentation style and information density influence the emergence and eventual resolution of *Vague Perception* phenomenon.

We construct benchmarks in which half of the samples are *verifiable* samples with *T/F* labels and the other half are *unverifiable* samples with *Unknown* labels that are objectively unsolvable. These samples are then fed into the model to extract the corresponding *Vague Perception* samples. As shown in Figure 1, LLMs sometimes output "*Unknown*" without a clear understanding of the sample. Subsequently, the selected ones are processed via a two-stage stimulation mechanism aimed at "waking up" the model's reasoning abilities. By measuring "Awaken Ability", we present a novel perspective on evaluating LLM reasoning, one that considers not only the model's initial accuracy, but also its potential to be guided for correction.

We propose WAKENLLM as a strong baseline to explore the reasoning abilities of LLMs among *Vague Perception* samples, yielding comparable performance across six different LLMs in these reasoning tasks, highlighting that both open-source and closed-source LLMs show substantial improvements in reasoning accuracy.

To summarize, our contributions are threefold:

- We dive into *Vague Perception*, a phenomenon in reasoning tasks that models output unknown because of subjective incapacity. To examine it comprehensively, we propose WAKENLLM, a unified framework that quantifies LLMs' capacity to convert to correct answers with stimulation—offering a principled solution for handling *Vague Perception* in reasoning tasks.
- Through systematic analyses across LLMs of varying parameter scales, we devise benchmark metrics that more precisely characterize the challenges posed by *Vague Perception*. The resulting insights not only boost reasoning accuracy but also deepen our understanding of LLM trustworthiness and advanced reasoning paradigms.
- By comparing WAKENLLM with diverse baselines, we measure how effectively each approach unlocks LLMs' potential. Incorporating both training and non-training frameworks, our work offers a new perspective on mitigating the *Vague Perception* phenomenon.

2 Evaluation Framework

The WAKENLLM evaluation framework, illustrated in Figure 2, is composed of two successive

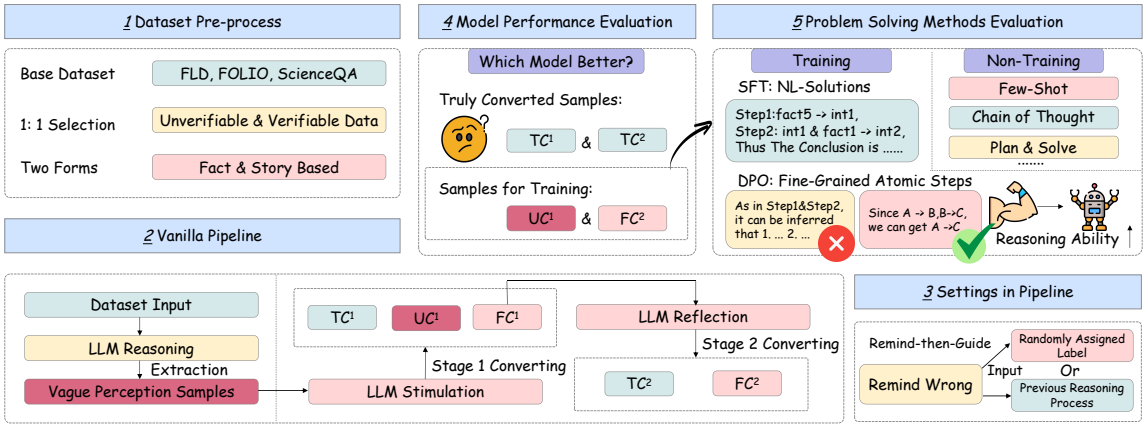

Figure 2: Overview of the proposed pipeline. We begin with a balanced Fact and Story-based dataset, sampling from verifiable (T/F) and unverifiable (Unknown) samples at 1 : 1 ratio. The pipeline then (1) detects cases that expose the model's *Vague Perception* during reasoning and (2) applies a two-stage remedy: Stage 1 uses stimulation to re-evaluate the mislabeled samples, while Stage 2 gathers the remaining *False Converting* samples from Stage 1 for model self-reflection. Stimulation Types vary with data form.

stages. It evaluates the model's latent reasoning ability and systematically evaluates the challenge of *Vague Perception*. For the vanilla setting, the first stage assesses the initial *Vague Perception* samples, while the second investigates the false converting samples FC^2, allowing us to compute the **Overall Benchmark Converting Rate**, indicating the percentage of *Vague Perception* that can be correctly reasoned by LLMs. Under the **Remind-then-Guide (RtG)** setting, in addition to reminding the model of its prior mistake, we either assign labels—true or false with a preset probability—or provide the previous reasoning process as input to examine whether the LLM still exhibits conformity. We examine whether LLMs still show conformity after being reminded that they solved the problems wrong previously, or it would learn from it and distinguish whether it is guiding or misguiding.

2.1 Overall Benchmark Converting Rate

Our framework begins with a model-specific preprocessing stage that identifies *Vague Perception* samples—cases where the LLM initially mislabels (e.g., saying *Unknown* when the sample is verifiable, or failing to explain an unverifiable sample). Since this process is performed separately for each model, the set of *Vague Perception* samples differs across each model.

After feeding the Stage 1 pipeline with *Vague Perception* samples, the conversion results fall into three categories: (1) The model fails to generate a coherent logic chain and remains output *Unknown*, which we call **Unexcited Converting (UC)**. (2) The model produces a false logic chain and an incorrect final label, denoted as **False Converting (FC)**. (3) The model outputs a correct label under stimulation, called **True Converting (TC)**. To quantify these proportions, we use **UCR**, **FCR**, and **TCR** to represent the rates of each type of conversion. Given our focus on the reasoning correction of the model, we elaborate only on TCR and OCR, defining **True Converting Rate(TCR)** in *n-th* stage as follows:

$$TC_f^n = \sum_{y_f \in D_{\text{VP}}^n} \mathbf{1}(\hat{y}_f^n = y_f), \quad TCR_f^n = \frac{|TC_f^n|}{|D_{\text{VP}}^n|}$$

In the above equations, $f \in \{u, v\}$ denotes the sample type (verifiable or unverifiable), and $n \in \{1, 2\}$ indicates the stimulation stage. D_{VP}^n represents the input of *Vague Perception* samples in the n-th stage. First stage input D_{VP}^1 was extracted from the reasoning dataset. For the second stage, we collect the **False Converting** part samples as the input, from which we can obtain: $D_{\text{VP}}^2 = FC^1$.

It should be noted that both TCR_v and TCR_u quantities represent the sum of results from stage 1 and 2, and TCR^1 denotes the converting rate of both forms data in first stage. To be specific, TCR_v = TCR_v^1 + TCR_v^2, and TCR^1 = TCR_v^1 + TCR_u^1. To integrate performance across both stages, we define **Overall Converting Rate (OCR)** as follows:

$$OCR = \frac{|TC^1 \cup TC^2|}{|D_{\text{vp}}|}, \quad TC_1 \cap TC_2 = \emptyset$$

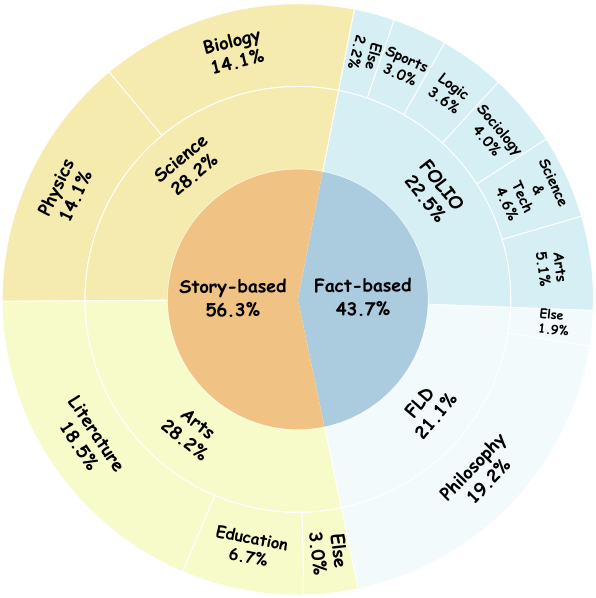

Figure 3: Overall benchmark distribution, which contains 2,840 samples grouped into two forms: **story-based** (56.3%) and **fact-based** (43.7%). A detailed breakdown of all topics and "Else" category is provided in the Appendix B.

OCR refers to the portion of under-stimulated samples that turn out to be predicted correctly. By doing this, we can obtain the latent accuracy of LLMs.

2.2 Remind-then-Guide (RtG) Conformity

Label Conformity We test whether *Vague Perception* samples display better or worse stability—i.e. whether they are more easily misguided—relative to ordinary settings indicated in (Weng et al., 2025). According to previous work, all models demonstrate **Conformity**. In this part, we first inform the model that it has misunderstood the sample, and then provide a randomly assigned label with varying probabilities, to test whether this misguiding affects its final decision.

We evaluate this phenomenon by setting up one metric for two types of samples: f can be v or u, and we split the results by whether the sample is verifiable or unverifiable. We compute the difference in accuracy between completely wrong and completely correct label guidance and then average the results. Detailed Remind-then-Guide conformity tests on assigning misguiding labels at four given possibilities (100%, 50%, 33.3%, 0%) is available in the Table 8, 9, 10, 11.

Reasoning Process (RP) Conformity We examine how effectively the reasoning chain generated in Stage 1 serves as inspiration or guidance for Stage 2 when it is supplied as additional input. To ensure that the model demonstrates genuine understanding rather than simply being prompted to output the correct answer by accident, we propose two metrics to dive deeper into this phenomenon.

Correct Guiding Rate (CGR) CGR assesses whether the reasoning process output in Stage 1 can assist Stage 2 in stimulating samples to better avoid prior mistakes. Similarly. We remind the LLMs that these explanations may contains flaws because the samples come from the **False Converting** set. CGR is calculated by the changes of TCR^2 with and without Reasoning Process.

Reasoning Process Coherence (RPC) This involves re-running Stage 1 with its own prior reasoning process to assess stability and coherence. Similarly, RPC is calculated by changes of TCR^1 with and without Reasoning Process.

For both metrics, RPC and CGR, below zero indicates that the previous reasoning process is misguiding and results in declined accuracy, while above zero suggests that it is helpful and can guide the LLM towards correct reasoning.

Table 2: Basic info of each dataset.

Dataset	Samples	Form	Writing	Source
FLD	600	**Conclusion**	Abstract	Gen
FOLIO	640	**Conclusion**	Plain	Real
Science	800	**Q&A**	Plain	Real
Arts	800	**Q&A**	Plain	Real

3 Experiments

3.1 Benchmarks

To explore the effects of the *Vague Perception* phenomenon on different types of logical datasets, we classify them based on their labels: (1) *Verifiable Data*, where the samples have a clear *True/False/Unknown* label indicating the validity of the reasoning; and (2) *Unverifiable Data*, where the label is *Unknown* and samples are unsolvable. Furthermore, as shown in Figure 3, to better understand how LLMs reason across different logical forms, we categorize benchmarks as follows:

Fact-based Benchmarks. The information is given in atomic relationships, split by unrelated sentences.: **FLD** (Morishita et al., 2024) supplies three kinds of sample labels annotated as *True*, *False*, or *Unknown*. We randomly sample 600 items: 300 verifiable (*True*/*False*) and 300 marked *Unknown* due to insufficient evidence. **FOLIO** (Han et al.,

Table 3: Comparison of six LLMs across four datasets (FLD, FOLIO, Science, Arts), grouped by evaluation settings. **RP** refers to Reasoning Process under Remind-then-Guide setting. The highest value is marked darker per column.

Model	Dataset	Vanilla			RtG (Label)		RtG (RP)	
		TCR[1]	TCR[2]	OCR	Conf_v	Conf_u	CGR	RPC
GPT-3.5	**FLD**	12.61	14.43	25.22	**64.25**	3.33	-12.30	10.84
	FOLIO	14.41	42.10	50.44	16.81	31.97	-42.10	-6.30
	Science	21.05	3.33	21.63	19.22	1.96	-3.33	0.59
	Arts	26.04	4.00	27.10	14.24	-3.88	0.00	3.12
GPT-4o	**FLD**	18.88	28.19	41.65	8.17	37.38	-23.50	-6.48
	FOLIO	37.34	44.40	65.16	9.63	51.90	-44.77	0.00
	Science	24.73	7.76	28.96	11.74	4.58	-7.16	-3.68
	Arts	18.60	8.24	24.17	8.07	23.31	-8.24	1.14
LLaMA-3.1-8B	**FLD**	41.97	26.03	48.90	50.04	11.04	-18.56	1.09
	FOLIO	45.86	11.40	50.84	38.73	13.61	-3.07	5.52
	Science	45.94	5.95	50.61	38.60	51.44	-5.95	12.91
	Arts	24.26	8.57	25.75	30.99	9.02	-4.77	11.30
LLaMA-3.1-405B	**FLD**	20.65	**53.73**	41.50	9.35	**57.63**	-48.28	-17.66
	FOLIO	28.76	29.92	49.22	6.11	23.01	-17.69	0.00
	Science	32.85	2.50	33.03	9.02	18.50	**0.19**	0.71
	Arts	4.40	0.00	4.40	0.81	28.91	0.00	-0.77
Qwen-2.5-7B	**FLD**	**48.31**	40.00	**68.53**	22.06	42.34	-31.67	-5.61
	FOLIO	19.73	36.84	47.36	4.92	38.71	-36.84	14.48
	Science	15.83	10.00	24.08	55.21	0.00	-2.98	**14.59**
	Arts	29.34	25.37	36.44	43.39	7.20	-13.27	8.47
Qwen-2.5-72B	**FLD**	48.20	9.80	51.21	33.08	9.14	-9.80	1.21
	FOLIO	31.51	11.37	38.36	10.65	2.20	-6.82	-1.37
	Science	10.81	3.12	11.20	33.42	22.07	0.00	-0.38
	Arts	25.00	9.76	25.56	12.52	24.12	0.00	-1.95

2022) is a factual benchmark with the same 3-way labeling. Similarly, we draw 640 samples—320 verifiable samples with *T/F* labels and 320 *Unknown*.

Story-based Benchmarks. Our experimental design considers the differing characteristics of the two forms above. For Fact-based ones, which models tend to process the entire context holistically, so providing additional stimulation can enhance performance. In contrast, we extract samples from ScienceQA (Lu et al., 2022) for story-based tasks—often rooted in professional knowledge domains and relying on models' inherent knowledge—we minimize guidance to avoid interfering with the model's reasoning process.

3.2 Experimental Setup

Our evaluation covers both *closed-source* models: ChatGPT-3.5-turbo and GPT-4o (OpenAI, 2023), and *open-source* models: LLaMA-3.1 series (Touvron et al., 2023), and Qwen-2.5 series (Bai et al., 2023). Two stimulating strategies are implemented:

▷ **Concise Stimulation**: Basic error notification without correction guidance.

▷ **Detailed Stimulation**: Enhanced detailed approach featuring:(1) Listed Conditions Guidance.(2) Structured Reasoning Constraints. (3) Conditional Verification Steps.The experimental protocol differs between approaches. *Concise Stimulation* uses a similar format for two-stage error notification process, while *Detailed Stimulation* employs a different two-stage workflow: a standard prediction followed by an interrogative refinement. The latter stimulates analytical reasoning through questioning and step-by-step error analysis, creating a clear methodological contrast.

To achieve a balance between randomness and coherence, we set all temperature to 0.5, except for 1.0 in ablation study where we want to further examine the quantitative influence between prompt designing and models setup.

In our result detection procedure, the output is first examined for specific words. When such methods are either absent or multiple candidates are detected, GPT-4o is employed to adjudicate and assign the appropriate label. With the exception of the **Remind-then-Guide (RtG)** setting in the reasoning process, most configurations yield a single label accompanied by few explanations, which facilitates reliable identification.

3.3 Results on Benchmark Converting Rate

As shown in Table 3, we report the overall benchmark Converting Rate, which means three metrics in three systems: True Converting Rate (TCR) for both stages, and Overall Converting Rate (OCR) taking the portion of *Vague Perception* samples into consideration. The LLaMA and Qwen series show superior Overall Converting Rate (OCR) on small-parameter series, whereas the GPT series shows

the opposite trend. We infer that, in general, higher parameter counts correlate with lower OCR. This situation may change due to the distinct yet controversial architecture of GPT-4o.

Table 4: Different Stimulating Method Influence on Converting Rate, where $\mathrm{TCR}^1 = \mathrm{TCR}^1_u + \mathrm{TCR}^1_v$ as we split converting rate by verifiable and unverifiable samples. *CS*, *DS* refer to *Concise* and *Detailed Stimulation*. Rows shaded gray denote ones used in our framework.

Metrics(%)	TCR^1	TCR^1_v	TCR^1_u	Reliability
FLD (CS)	37.27	10.65	26.62	↓
FLD (DS)	18.88	**13.64**	5.24	↑
Arts (CS)	18.60	**16.06**	2.54	↑
Arts (DS)	22.54	14.65	7.89	↓

3.4 Results on RtG Conformity Stimulating

As illustrated in the Figure 2, RtG module contains two parts: Label and Reasoning Process (RP) Conformity: We remind the model incorrectly predicted the sample as *Unknown* before; we then we apply conformity test to verify whether models can inference independently.

Label Conformity Judging from the table, we observe that even when LLMs are reminded of their previous incorrect answers, they still demonstrate a strong tendency toward conformity. Moreover, distinct conformity patterns emerge between verifiable and unverifiable samples. One key influencing factor appears to be the model size: models with smaller parameter sizes, such as GPT-3.5 and LLaMA-3.1-8B-Instruct, are more inclined to being misguided by verifiable samples with *T/F* labels reasoning. In contrast, models with larger parameter sizes, such as GPT-4o and LLaMA-3.1-405B, tend to be more easily misguided regarding unverifiable samples with *Unknown label*. Meanwhile, Qwen series shows distinct yet related features compared with the above, we infer it is because parameters gap between Qwen 7B and 72B is much smaller than the former two series.

Reasoning Process (RP) Conformity We input the previous reasoning process and explicitly ask the LLM to reflect on its mistakes, the resulting CGR values are mostly below zero, which indicates that even after being reminded of a past error and provided with its own flawed reasoning, the LLM still tends to repeat the same incorrect conclusion, leading to an even larger decline in accuracy.

Our results on RPC reveal that the reasoning process is often highly unstable. Changes in accuracy can largely be attributed to divergence in how the problem is interpreted. Notably, for unverifiable samples, the reasoning process tends to be unreliable — feeding it back into the model typically leads to a further drop in accuracy. This exposes a hidden issue: when LLMs automatically generate solutions for a dataset using a reasoning process, unverifiable problems are particularly susceptible to unreliability. Even for verifiable samples, the reasoning may remain unstable. As a result, **the accuracy and reliability of AI-annotated problem-solving methods for datasets remain a problem.**

4 Detailed Analysis

4.1 True Converting Rate (TCR) Analysis

As shown in the line chart of Figure 4, by comparing TCR^1 and TCR^2 across different benchmarks, we observe clear divergences: This indicates that many samples mis-predicted in Stage 1 can be self-corrected in Stage 2, where the model refines its initial output. In contrast, Story-based TCR^1 shows a substantially lower TCR^2. Whereas LLaMA-3.1-8B demonstrates much higher TCR^1.

Thus, distinct focus strategies are recommended for multi-stage frameworks for the following LLMs: (i) For GPT-4o, when facing *Fact-based* benchmarks, broader exploratory freedom is acceptable in Stage 1. “making mistakes at first is acceptable", since later refinement proves beneficial; For *Story-based* benchmarks, the improvement of the subsequent stage is limited, so initial accuracy should be emphasized (ii) For LLaMA-3.1-8B, it is always better to “do it right the first time" .

4.2 Benchmark-wise Vague Perception

As each model differs in its reasoning ability, we first quantify the percentage of data in which models induce *Vague Perception*. The Bar chart in blue Fig 4 reports the percentage for GPT-4o—the most widely used LLM at present—while complete results for the remaining models in Appendix 7.

We can conclude that **the Vague Perception phenomenon is prevalent across all benchmarks**. For GPT-4o, FOLIO exhibits a relatively lower percentage. For FLD and Arts—which focus purely on language understanding, the model demonstrates greater uncertainty regarding explaining why certain samples are labeled as *Unknown*; In stark contrast, *Science*—which comprises Physics and Biol-

ogy topics—faces significant challenges even when the samples are verifiable; For LLaMA-3.1-8B, there is a significant rise in the percentage of verifiable samples predicted as *Unknown*, but the ability to comprehend unverifiable samples is stronger.

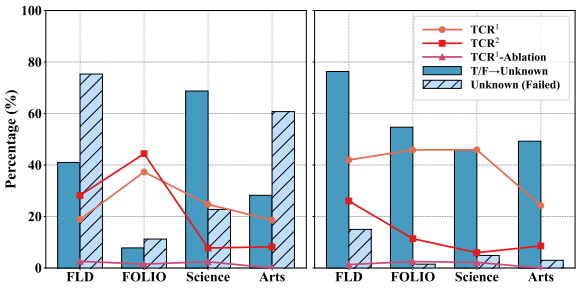

Figure 4: Percentage of GPT-4o (left) and LLaMA-3.1-8B (Right) that Exhibit *Vague Perception* Across Different benchmarks. "T/F→Unknown" indicates that are verifiable in principle but which the model incorrectly predicted as "*Unknown*" . "Unknown (Failed)" denotes cases where the model outputs "*Unknown*" because it lacks the capability to correctly assess verifiability.

4.3 Stimulating Style Comparison

To evaluate the effectiveness, we swapped the two styles of stimulation within our framework and ran the experiment, using GPT-4o as the base model. As shown in Table 4, **switching stimulation styles leads to a drop in TCR^1_v**, indicating weakened understanding on verifiable *T/F* samples. In contrast, TCR^1_u increases for unverifiable *Unknown* samples. Although the overall conversion improves, this gain is unstable, as TCR^1_v remains more reliable.

Furthermore, we observe a latent trade-off: **as the ability to resolve verifiable samples declines, the model becomes more confident on unverifiable ones**. We attribute this to a shift in the model's reasoning habits.

4.4 Ablation Study

From the Stage 1 output, we isolated the Unexcited samples—those that failed to even generate a complete logic chain—and fed them back through the same Stage 1 prompting protocol. Both LLaMA-3.1-8b-Instruct and GPT-4o are set with 1.0 temperature to ensure randomness.

As Figure 4 shows, this second pass produces virtually no additional conversions: more than 95% of the samples remain unchanged. This finding demonstrates that *Vague Perception* is largely immune to repeated, identical n-shot prompting; applying the same stimulation twice is therefore futile. **Consequently, any effective correction must rely on a distinct stage 2 rather than reiterating stage 1, adjusting temperature alone without revising stimulation style can be limited.**

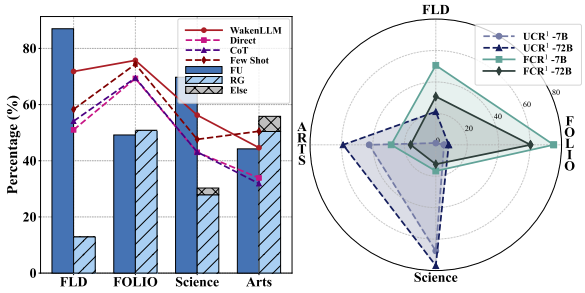

Figure 5: Distribution of the root causes behind GPT-4o's *Vague Perception* errors across all benchmarks (left). Qwen-2.5-7B and 72B model False and Unexcited Converting Rate comparison (Right). "FU": *Fact Understanding*, "RG": *Reasoning Gap*, and "Else" aggregates residual minor factors.

4.5 Baseline Comparison

In our work, we primarily analyze the "waken ability" with respect to *Vague Perception* samples by computing the proportion of truly converted samples in both stages (TC^1 and TC^2) relative to the percentage of *Vague Perception* samples; however, this constitutes only one component, we combine this component with the vanilla accuracy by adding them together, thus we gain an empirical bound of reasoning accuracy—namely, the maximum reasoning capability that LLMs can achieve without additional training or data-processing. The empirical accuracy of WAKENLLM is defined by two parts: (1) the proportion of samples correctly reasoned by the LLM, and (2) the truly converted samples across both stages (TC^1, TC^2) that are potentially solvable. *Acc(Vanilla)* denotes the reasoning accuracy under direct prompt setting, without any intermediate interpretation. Here, D represents the complete dataset, whereas D_{VP} only represents a shortlisted ones that induces *Vague Perception*.

$$\text{Acc(WAKENLLM)} = \underbrace{\text{Acc(Vanilla)}}_{\text{Explicit Ability}} + \underbrace{|TC^1 \cup TC^2|/|D|}_{\text{Latent Ability}}$$

As illustrated in the left section of the line chart in Figure 5, most baseline methods exhibit a substantial reasoning gap compared to our pipeline, indicating most current frameworks failed to trigger models' full potential. Refer to the Table 6 for detailed experimental results.

Table 5: Performance comparison of Degradation Analysis Phenomenon between GPT-4o and GPT-3.5.

Models	**GPT-4o**		**GPT-3.5**	
Metrics(%)	**TCR^2**	**Deg**	**TCR^2**	**Deg**
FLD	28.19↑	3.08	14.43↓	3.09
FOLIO	44.40↑	5.22	42.10↓	6.32
Science	7.76↑	**26.86**	3.33↓	3.33
Arts	8.24↑	1.13	4.00↓	0.00

4.6 Phenomenon Exploration

To probe the root causes of GPT-4o's mistakes, we annotate each error sample with one of three labels:
▷ **Fact Understanding (FU)** – The model does not correctly retrieve, align, or synthesize the relevant facts stated in the passage.
▷ **Reasoning Gap (RG)** – The model retrieves the facts but fails to perform the intermediate logical operations that connect facts to the hypothesis.
▷ **Else** – For the two above generally covers 80% or more scenarios. And there is a catch-all category for other less frequent factors. Notably, issues related to **Excessive Caution** are particularly prominent in this part.

Excessive Caution Our experiments reveal that Excessive Caution—instances where the model responds with *Unknown* despite sufficient evidence—is nearly absent in *Fact-based* benchmarks but increases from 2.4% in **Science** to 5.4% in **Arts**. We attribute this rise to the narrative complexity of *Story-based* tasks: unlike the atomic and structured format of *Fact-based* benchmark, passages in **Arts** often embed literary allusions and background knowledge, heightening epistemic uncertainty.

4.7 Degradation Analysis

Within the pipeline, we don't exclude the possibility that previous incorrectly predicted samples, FC^2 from Stage 1 may be converted to Unknown, which should be denoted as UC^2. As it is reasonable that at first, the model may generate an incorrect answer, but after further stimulation, due to its limited ability, it fails to arrive at the correct answer and can only output *Unknown*. We elaborate this as:

$$Deg = \frac{\left|FC^1 \setminus \left(FC^2 \cup TC^2\right)\right|}{|FC^1|},\ FC^2 \cap TC^2 = \emptyset$$

As shown in Table 5, only a small proportion of samples exhibit this phenomenon. However, in the case of GPT-4o on the reasoning-based Science dataset, the issue becomes more serious.

As for the **Science dataset**, GPT-4o achieves the highest TCR^2 among all four datasets. Coupled with higher *Deg* value, this suggests that **when confronted with problems it previously answered incorrectly—especially those beyond its capability—GPT-4o tends to respond with an *Unknown* label. In contrast, GPT-3.5-turbo is much more likely to provide a definitive but incorrect answer instead of admitting uncertainty.**

This can be further proved by the uniform distribution of UCR and FCR in Appendix D acquired similar to formula TCR, denoting Unknown and False Converted rate in the *Vague Perception* input. We compare answering of Qwen-2.5 series (7B and 72B model), as shown in the radar-chart in Figure 5: (1) Models show distinctive yet uniform preferences among four different forms of dataset. For FLD and FOLIO, **models are much more willing to output a false answer rather than admitting Unknown**, this is even more serious for FOLIO dataset, where model likes to "show off" when facing easier ones, as FOLIO achieves the highest accuracy among above four; **Story-based dataset**, like Science and Arts are **exactly the opposite**. (2) The higher the parameters, the more likely models are to output *Unknown* when inputting problems beyond LLMs' ability.

5 Conclusion

We present WAKENLLM, a fine-grained framework for evaluating the *Vague Perception* phenomenon of LLMs. The framework supplies targeted cues that stimulate models to reflect on reasoning, substantially raising the empirical accuracy. Experiments show that while LLMs give up and answer *Unknown* on the first try, well-designed stimulation reliably steers them to the correct label and uncovers a persistent tendency toward conformity even after explicit reminding. These insights suggest training and evaluation protocols that value a model's capacity for potential instead of the one-shot performance alone. Even though significant improvements in inner structures and parameter optimizations of LLMs may be limited in the near future, however, our findings suggest that current methods are still far from triggering LLMs' full potential, and addressing this problem can bring unique prospective. Our research offers high underlying benefits regarding models' potential.

6 Limitations

Even though our work has helped expose the empirical reasoning potential of LLMs, however, our final metrics are all calculated regarding labels, and outputting true labels with flawed reasoning process can be a common problem. How to fundamentally quantify their true ability with true reasoning process remains a problem in the future.

7 Ethical Considerations

Our research utilizes publicly available datasets (FLD, FOLIO, ScienceQA) containing synthetic logical reasoning problems without personal information. We acknowledge potential biases in our SFT and DPO training processes and implement atomic reasoning criteria to ensure fair evaluation. All training scripts, evaluation code, and prompt templates are made publicly available for reproducibility. We use LoRA fine-tuning to reduce computational requirements and focus on improving logical reasoning quality rather than enabling harmful applications. Our work adheres to academic integrity standards and contributes to responsible AI development through more interpretable reasoning.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Sébastien et al. Bubeck. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Mingyang Chen, Tianpeng Li, Haoze Sun, Yijie Zhou, and 1 others. 2025. Research: Learning to reason with search for llms via reinforcement learning. *arXiv preprint arXiv:2503.19470*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, and 1 others. 2022. Palm: Scaling language modeling with pathways. In *Proceedings of ICML*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2924–2936. Association for Computational Linguistics.

Karl Cobbe and 1 others. 2021. Training a helpful and harmless assistant with reinforcement learning. *arXiv preprint arXiv:2104.08718*.

Clayton Cohn, Nicole Hutchins, Ashwin T. S., and Gautam Biswas. 2025. Cotal: Human-in-the-loop prompt engineering, chain-of-thought reasoning, and active learning for generalizable formative assessment scoring. *Preprint*, arXiv:2504.02323.

André de Souza Loureiro, Jorge Valverde-Rebaza, Julieta Noguez, David Escarcega, and Ricardo Marcacini. 2025. Advancing multi-step mathematical reasoning in large language models through multi-layered self-reflection with auto-prompting. *arXiv preprint arXiv:2506.23888*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xiang Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3563–3578. Association for Computational Linguistics.

Shizhe Diao, Pei Wang, Yujia Lin, Ruili Pan, Xuefeng Liu, and Tong Zhang. 2024. Active prompting with chain-of-thought for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, pages 1330–1350. Association for Computational Linguistics.

Shizhe Diao, Pengcheng Wang, Yong Lin, Rui Pan, Xiang Liu, and Tong Zhang. 2023. Active prompting with chain-of-thought for large language models. arXiv preprint arXiv:2302.12246. Cs.CL.

Yilun Du, Sherry Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, pages 11733–11763. PMLR.

Luyu Gao, Aman Madaan, Shuning Zhou, Uri Alon, Pengfei Liu, Yinhan Yang, Jamie Callan, and Graham Neubig. 2022. PAL: Program-aided language models. *arXiv preprint arXiv:2211.10435*.

Zhipeng Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yang Yang, Nan Duan, and Weizhu Chen. 2024. Critic: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations*.

Thilo Hagendorff and Stefan Fabi. 2025. Beyond chains of thought: Benchmarking latent-space reasoning abilities in large language models. *arXiv*, 2501.01234. https://arxiv.org/abs/2501.01234.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, and 1 others. 2022. Folio: Natural language reasoning with first-order logic. *arXiv preprint arXiv:2209.00840*.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. 2023. Reasoning via planning: Language model is planning with world model. *arXiv preprint arXiv:2305.14992*.

Dan Hendrycks and et al. 2021. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Boyi Hou, Yucheng Liu, Kun Qian, Jacob Andreas, Shiyu Chang, and Yiming Zhang. 2024. Decomposing uncertainty for large language models through input clarification ensembling. In *Proceedings of the 41st International Conference on Machine Learning*, pages 19023–19042. PMLR.

Xinyan Huang, Shuo Li, Mo Yu, Maurizio Sesia, Hadi Hassani, Insu Lee, Osbert Bastani, and Edgar Dobriban. 2024. Uncertainty in language models: Assessment through rank-calibration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 284–312. Association for Computational Linguistics.

Chujie Jiang, Runzhe Dong, Xin Liu, Diyi Xiong, and Yiming Yang. 2024. Honestllm: Measuring and improving truthfulness in large language models. In *Findings of the Association for Computational Linguistics*, pages 1234–1250.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. Decomposed prompting: A modular approach for solving complex tasks. *Preprint*, arXiv:2210.02406.

Tatsunori Kojima, Shixiang Shane Gu, Mike Reid, Yutaka Matsuo, Yutaka Iwasawa, and Haruo Kamei. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3835–3849.

Aitor Lewkowycz and et al. 2022. Solving quantitative reasoning problems with language models. In *arXiv preprint arXiv:2206.14858*.

Ziqing et al. Lin. 2023. Exposing smart chat bots via multiple personas red teaming. *arXiv preprint arXiv:2305.17414*.

Zipeng Ling, Yuehao Tang, Chen Huang, Shuliang Liu, Gaoyang Jiang, Shenghong Fu, Junqi Yang, Yao Wan, Jiawan Zhang, Kejia Huang, and Xuming Hu. 2025. Instruction boundary: Quantifying biases in llm reasoning under various coverage. *Preprint*, arXiv:2509.20278.

Chen Liu and 1 others. 2025. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models. *arXiv preprint arXiv:2503.04567*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. arXiv preprint arXiv:2301.13379. Cs.AI.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Peter Clark, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*.

Tyler McDonald and Ali Emami. 2025. Trace-of-thought prompting: Investigating prompt-based knowledge distillation through question decomposition. *arXiv preprint arXiv:2504.20946*.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. Ambigqa: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*. Association for Computational Linguistics.

Terufumi Morishita, Gaku Morio, Atsuki Yamaguchi, and Yasuhiro Sogawa. 2024. Enhancing reasoning capabilities of llms via principled synthetic logic corpus. *Advances in Neural Information Processing Systems*, 37:73572–73604.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, and et al. 2021. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*.

Tobias Olausson, Albert Gu, Brandon Lipkin, Chiyuan Zhang, Armando Solar-Lezama, Joshua Tenenbaum, and Roger Levy. 2023. Linc: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5153–5176. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Long Ouyang, Jeff Wu, Xu Jiang, and *et al.* 2022. Training language models to follow instructions with human feedback. *arXiv:2203.02155*.

Rafael Rafailov, Usama Chughtai, Yining Zhang, Xuechen Li, Colin Raffel, and Tatsunori Hashimoto. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*. Association for Computational Linguistics.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint arXiv:2303.11366*.

Aarohi Srivastava and et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. In *International Conference on Learning Representations*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: A large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2018)*. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Bohan Wang, Xiang Zhang, Cheng Liu, and Pan Zhou. 2023a. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. *arXiv preprint arXiv:2305.04091*.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023b. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. *arXiv preprint arXiv:2305.04091*.

Wei Wang and Yang Song. 2024. Mars: Benchmarking the metaphysical reasoning abilities of language models with a multi-task evaluation dataset. *arXiv preprint arXiv:2407.00888*.

Xuezhi Wang, Ethan Zhou, Abulhair Saparov, Quoc Le, and Hsien-Yu Liu. 2023c. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, and et al. 2022a. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Zhou, Jay Li, Quoc Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Zhiyuan Weng, Guikun Chen, and Wenguan Wang. 2025. Do as we do, not as you think: the conformity of large language models. *Preprint*, arXiv:2501.13381.

Zhenyu Wu, Qingkai Zeng, Zhihan Zhang, Zhaoxuan Tan, Chao Shen, and Meng Jiang. 2024. Large language models can self-correct with key condition verification. *Preprint*, arXiv:2405.14092.

Jing Xu, Hengrui Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. 2024. Faithful logical reasoning via symbolic chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, pages 13326–13365. Association for Computational Linguistics.

Silei Xu, Wenhao Xie, Lingxiao Zhao, and Pengcheng He. 2025. Chain of draft: Thinking faster by writing less. *arXiv preprint arXiv:2502.18600*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*. Association for Computational Linguistics.

Shinn Yao, Jeffrey Zhao, Dian Yu, Sherry Zhao, Nathan Wang, Bill Wang, and Denny Zhou. 2023a. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Shunyu Yao, Shiyu Liang, Jeffrey Zhao, and Karthik Narasimhan. 2023b. Tree of thoughts: Deliberate reasoning via tree search. *arXiv preprint arXiv:2305.10601*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Yunxiang Zhang, Muhammad Khalifa, Lajanugen Logeswaran, Jaekyeom Kim, Moontae Lee, Honglak Lee, and Lu Wang. 2024. Small language models need strong verifiers to self-correct reasoning. *Preprint*, arXiv:2404.17140.

Jiaxing Zhao, Hongbin Xie, Yuzhen Lei, Xuan Song, Zhuoran Shi, Lianxin Li, Shuangxue Liu, and Haoran Zhang. 2025. Connecting the dots: A chain-of-collaboration prompting framework for llm agents. *Preprint*, arXiv:2505.10936.

Shunyu Zhou, Surya Pal, Silvio Savarese, and Amir Zamir. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

Yujun Zhou, Jiayi Ye, Zipeng Ling, Yufei Han, Yue Huang, Haomin Zhuang, Zhenwen Liang, Kehan Guo, Taicheng Guo, Xiangqi Wang, and Xiangliang Zhang. 2025. Dissecting logical reasoning in llms:

A fine-grained evaluation and supervision study. *Preprint*, arXiv:2506.04810.

Dong-Hai Zhu, Yu-Jie Xiong, Jia-Chen Zhang, Xi-Jiong Xie, and Chun-Ming Xia. 2025. Understanding before reasoning: Enhancing chain-of-thought with iterative summarization pre-prompting. *Preprint*, arXiv:2501.04341.

A Baseline Details

Chain-of-Thought (CoT) The CoT prompting method guides large language models (LLMs) to generate intermediate reasoning steps before producing a final answer. It improves multi-step reasoning performance by making the inference process explicit. This technique is particularly effective in arithmetic and logic-heavy tasks, as shown in prior works such as Wei et al. (2022b).

Few-shot Few-shot prompting involves presenting the model with a few annotated examples in the input to demonstrate the task format. It enables in-context learning without any gradient updates. Despite its simplicity, few-shot prompting serves as a strong baseline in many LLM applications (Brown et al., 2020).

Tree-of-Thought (ToT) Tree-of-Thought extends the CoT framework by exploring multiple reasoning paths in a tree structure and evaluating intermediate steps. It mimics a planning process that considers alternative possibilities, significantly enhancing reasoning robustness and allowing backtracking or reevaluation of incorrect paths (Yao et al., 2023a).

Plan & Solve Plan-and-Solve separates the reasoning process into two stages: first planning a high-level outline or decomposition of the task, then solving sub-tasks based on that plan. This two-stage strategy enhances controllability and consistency in reasoning, and has shown strong performance across complex benchmarks (Wang et al., 2023a).

Supervised Fine-Tuning (SFT) Supervised Fine-Tuning (SFT) is the standard method for aligning LLMs by training on curated human-annotated question-answer pairs. Unlike prompt-only methods, SFT optimizes the model weights directly through gradient updates to produce desirable outputs. It forms the foundation of instruction-tuned models like InstructGPT and serves as a widely-used supervised baseline for alignment tasks (Ouyang et al., 2022). The common consensus is that SFT provides the essential starting point for alignment, but by itself is insufficient to fully capture complex human preferences, often requiring further preference optimization or reinforcement learning. The optimized formula is below:

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\sum_{i=1}^{N} \log \pi_\theta(y_i \mid x_i),$$

Direct Preference Optimization (DPO) Direct Preference Optimization (DPO) is a reinforcement learning-free approach that fine-tunes LLMs using preference data. Given pairs of outputs (e.g., preferred vs. dispreferred), DPO directly optimizes the log-likelihood ratio between them, providing alignment benefits similar to reinforcement learning methods like PPO while avoiding their instability and complexity (Rafailov et al., 2023). The common consensus is that DPO offers a simpler and more stable alternative to reinforcement learning-based methods, making it a practical choice for preference alignment, though it may still struggle with capturing nuanced or long-term user preferences. The formula can be concluded as below:

$$\mathcal{L}_{\mathrm{DPO}}(\theta) = -\mathbb{E}_{(x,y^+,y^-)}\Big[\log \sigma\Big(\beta\big(\log \tfrac{\pi_\theta(y^+|x)}{\pi_{\mathrm{ref}}(y^+|x)} - \log \tfrac{\pi_\theta(y^-|x)}{\pi_{\mathrm{ref}}(y^-|x)}\big)\Big)\Big]$$

A.1 Experiment Setup

We choose four LLMs: GPT-4o, GPT-3.5-turbo-1105, LLaMA-3.1-8B and Qwen-2.5-7B for our detailed baseline experiments. We categorize our baseline methods into two kinds: training and non-training, in order to see whether the *Vague Perception* phenomenon can be solved with or without parameter revision. For model training and open-source model deployment, the experiments ran on four Tesla V100-SXM2-32GB GPUs.

A.2 Training Methods Details

As explained before in Appendix A.1, we elaborate our facilities and setup for training. In this section, we demonstrate our details for the training.

For SFT (Supervised Fine-Tuning), we extract 1,000 training samples from exactly same dataset, the prompt instruction can be found in Appendix D.2.

For DPO (Direct Preference Optimization), we apply LLM Judge to help select the reasoning process with many atomic steps as possible. This includes two parts: (1) Prompt Template that includes

task guiding details (e.g., definition of atom steps) and task instructions (2) Testing cases that can perfectly prove and verify which response is better given the metrics and requirements. All information above can be found in Appendix D.2.

B Benchmark Details

FLD Following Morishita et al. (2024), **FLD** is a fully synthetic benchmark for multi-step logical inference. Formal proof chains of depth 0–15 are automatically generated and then verbalized into natural-language premises plus a hypothesis. Each instance is labeled **Proved**, **Disproved**, or **Unknown**, making the dataset ideal for assessing whether a model can (i) draw correct conclusions and (ii) abstain when evidence is insufficient.

FOLIO **FOLIO** (Han et al., 2022) contains **1,430** human-written first-order-logic stories that bridge everyday commonsense with formal theorem proving. Every narrative provides its gold logical forms and a hypothesis whose truth value (**True**, **False**, or **Unknown**) is automatically certified. The official split is 1 001/226/203 for train/dev/test, offering fewer but richer problems than FLD and requiring explicit refusals when the passage leaves the hypothesis undecided.

ScienceQA **ScienceQA** (Lu et al., 2022) is a multi-modal multiple-choice benchmark with **21,208** elementary- and middle-school questions across 26 topics. To create an abstention setting comparable to the logic corpora, we select questions from Physics and Biology (science) as well as Writing-Strategies and Figurative-Language (arts). For 50% of the chosen items we delete two random supporting sentences with GPT-4o, then manually verify that no unique answer remains; these items are re-labeled **UNKNOWN**. The final corpus therefore mixes standard A–D supervision with an explicit refusal class.

Topic Analysis Our comprehensive dataset consists of four distinct collections, each representing different domains of logical reasoning. The **FLD (First-Order Logic Dataset)** contains 600 samples across 11 topics, with Philosophy being the dominant subject (545 samples, 90.8%), followed by Logic (37 samples, 6.2%), and 9 other topics with minimal representation. The **FOLIO dataset** comprises 640 samples distributed across 7 topics, including Arts (146 samples, 22.8%), Science & Technology (130 samples, 20.3%), Sociology (114 samples, 17.8%), Logic (102 samples, 15.9%), Sports (85 samples, 13.3%), Geography (39 samples, 6.1%), and Architecture (24 samples, 3.8%). The **ScienceQA dataset** contains 800 samples evenly split between Biology and Physics (400 samples each, 50%). Finally, the **Arts dataset** includes 800 samples with Literature as the primary topic (526 samples, 65.8%), followed by Education (190 samples, 23.8%), Philosophy (77 samples, 9.6%), Linguistics (4 samples, 0.5%), and Psychology (3 samples, 0.4%).

The dataset distribution reveals a hierarchical structure: at the top level, we distinguish between **Fact-based reasoning** (43.66%) and **Story-based reasoning** (56.34%). The four major categories are distributed as follows: FLD (21.13%), FOLIO (22.53%), Science (28.17%), and Arts (28.17%). This balanced distribution ensures comprehensive coverage of different reasoning paradigms, from formal logical reasoning in FLD to narrative-based reasoning in Arts, providing a robust foundation for training and evaluating logical reasoning capabilities across diverse domains.

Unknown Labeling For the first **FLD** and **FOLIO**, the *Unknown* label is originally included in their annotations. For SCIENCEQA, we randomly select half of the samples and employ GPT-4o to randomly delete two sentences from the given facts. We then re-label these modified samples as *Unknown*. Additionally, we perform manual verification to ensure that each relabeled sample is semantically consistent with its assigned label.

Ethical Statement All datasets are distributed for non-commercial research under permissive licenses (FLD: MIT; FOLIO: research-only; SCIENCEQA: CC-BY-NC-SA 4.0). Our sentence-removal augmentation introduces no personal data and was human-verified to avoid unsafe or misleading content.

Search-Based Planning Frameworks Instead of generating a single chain, **Tree-of-Thought (ToT)** views partial solutions as nodes in a tree and performs breadth-first search with value heuristics (Yao et al., 2023b). **Decomposed Prompting** turns a complex task into callable sub-prompts that can be solved (and verified) independently, achieving strong zero-shot gains on compositional reasoning benchmarks (Khot et al., 2023). *Reasoning-via-Planning (RAP)* couples an LLM world-model with Monte-Carlo Tree Search for

Model	Methods	Accuracy (%)			
		FLD	FOLIO	Science	Arts
GPT-3.5	Direct	50.17	40.63	44.63	29.25
	CoT	49.67	42.34	44.13	31.00
	Few-shot	49.67	37.50	43.75	33.88
	Tree-of-Thought	33.50	43.91	36.00	28.25
	Plan & Solve	42.83	45.47	33.63	26.38
	WAKENLLM	**60.26**	**46.68**	**48.38**	**35.04**
GPT-4o	Direct	51.00	69.53	43.00	33.88
	CoT	54.10	69.53	43.12	31.87
	Few-shot	58.30	74.40	47.62	**50.50**
	Tree-of-Thought	52.83	70.62	43.38	29.25
	Plan & Solve	52.17	69.84	42.38	30.25
	WAKENLLM	**71.79**	**75.74**	**56.25**	44.64
LLaMA-3.1-8B	Direct	39.17	51.56	49.50	47.50
	CoT	40.67	51.88	46.50	43.25
	Few-shot	33.33	51.25	48.25	39.13
	Tree-of-Thought	38.33	47.66	40.63	30.00
	Plan & Solve	35.50	52.81	38.75	26.75
	SFT	49.17	50.47	46.50	49.12
	DPO	50.33	48.60	48.50	49.62
	WAKENLLM	**61.50**	**65.47**	**62.72**	**53.43**
Qwen-2.5-7B	Direct	41.83	57.81	44.12	47.12
	CoT	43.83	60.62	42.75	46.62
	Few-shot	44.50	58.75	38.00	42.25
	Tree-of-Thought	39.83	57.19	37.88	45.88
	Plan & Solve	40.67	56.88	37.62	39.62
	SFT	48.67	57.34	**49.62**	43.12
	DPO	**49.33**	54.21	48.38	44.13
	WAKENLLM	48.06	**65.55**	47.56	**54.37**

Table 6: Accuracy comparison of multiple baselines for four LLMs evaluated on four datasets. Each block represents a distinct model series. The highest accuracy is marked black.

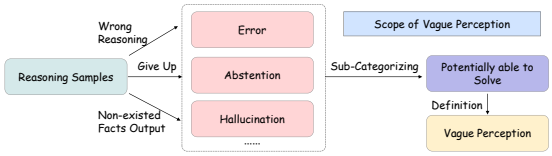

Figure 6: Scope of *Vague Perception*. This phenomenon underlies a wide range of reasoning failures. By identifying and sub-categorizing cases that are potentially solvable, it reveals a substantial portion of failures that stem from *Vague Perception* rather than fundamental incapacity. Addressing this issue can significantly advance fine-grained reasoning and improve the overall trustworthiness of LLMs.

long-horizon tasks (Hao et al., 2023). **Cochain** (*Chain-of-Collaboration Prompting*) extends this line by orchestrating multiple lightweight agents whose specialized sub-chains are merged via a knowledge-graph plan, outperforming single-agent CoT on workflow-style tasks while using fewer tokens (Zhao et al., 2025).

C Related Work Details

Chain-of-Thought (CoT) prompting (Wei et al., 2022a) first showed that furnishing step-wise exemplars markedly improves LLM performance on arithmetic and commonsense problems. Follow-up work refines this paradigm: *scratchpads* (Nye et al., 2021), *least-to-most prompting* (Zhou et al., 2022), and *self-consistency* (Wang et al., 2023c). More recently, multi-agent and symbolic reasoning approaches like *multiagent debate* (Du et al., 2024), *symbolic chain-of-thought* (Xu et al., 2024), and neurosymbolic reasoning with logic provers (Olausson et al., 2023) further strengthen factuality and logical consistency.

Tool-Augmented Reasoning **Program-Aided LMs** (Gao et al., 2022) delegate arithmetic or logical computation to Python. **ReAct** (Yao et al., 2022) interleaves reasoning and external tools. *CRITIC* (Gou et al., 2024) validates interme-

Model	FLD		FOLIO		Science		Arts	
	V	U	V	U	V	U	V	U
GPT-3.5	72.00	8.00	10.00	24.69	40.00	2.75	22.50	1.50
GPT-4o	11.00	88.00	7.81	11.25	68.75	22.75	28.25	60.75
LLaMA-3.1 (8B)	76.33	15.00	54.69	1.88	49.25	3.00	46.00	4.50
LLaMA-3.1 (405B)	22.67	100.00	20.94	99.69	50.00	89.25	9.00	99.00
Qwen-2.5 (7B)	20.67	9.00	10.00	13.75	60.00	0.00	52.50	6.50
Qwen-2.5 (72B)	49.67	5.67	15.31	7.50	54.75	10.00	27.75	36.25

Table 7: The percentage of LLMs demonstrating the *Vague Perception* phenomenon among six LLMs. Split by two different root causes: (1) V: Verifiable samples with *T/F* labels, but the model output *Unknown*. (2) U: Unverifiable Samples with *Unknown* label, but the model output *Unknown* simply because of incapacity rather than fully understand this is objectively unsolvable.

diate answers with execution tools. *Chain-of-Verification* (Dhuliawala et al., 2024) reduces hallucinations by verifying sub-claims against evidence.

Iterative Self-Refinement and Critique Self-Refine (Madaan et al., 2023) establishes a generate–criticise–rewrite loop. Ambiguity handling and uncertainty decomposition are also explored: *AmbigQA* (Min et al., 2020), *Decomposing Uncertainty* (Hou et al., 2024), and rank-based calibration (Huang et al., 2024).

Datasets for Reasoning and Verification Classic benchmarks like *BoolQ* (Clark et al., 2019), *SQuAD 2.0* (Rajpurkar et al., 2018), *FEVER* (Thorne et al., 2018), and *HotpotQA* (Yang et al., 2018) stress-test fact verification, ambiguity, and multi-hop reasoning.

Active Prompting Active-Prompt (Diao et al., 2024) selects informative exemplars to boost performance under budget constraints. Meanwhile, Ling et al. (2025) denotes different prompt coverage leads to various effects.

Faithfulness and Safety Beyond correctness, methods ensure interpretability and reliability, such as *Faithful-CoT* (Lyu et al., 2023), symbolic verification (Xu et al., 2024), and multi-agent debate (Du et al., 2024).

D Formula Details

D.1 False and Unknown Converting

As demonstrated in 2.1, we denote the notion of **TCR**. Likewise, we define the concepts regarding false converting and unknown converting similarly:

$$FC_f^n = \sum_{y_f \in D_{\text{VP}}^n} \mathbf{1}\big(\hat{y}_f^n \neq y_f,\ \hat{y}_f \notin U\big)$$

$$FCR_f^n = \frac{|FC_f^n|}{|D_{\text{VP}}^n|}$$

$$UC_f^n = \sum_{y_f \in D_{\text{VP}}^n} \mathbf{1}\big(\hat{y}_f^n \in U,\ y_f \notin U\big)$$

$$UCR_f^n = \frac{|UC_f^n|}{|D_{\text{VP}}^n|}$$

Note that, Unknown Converting (UC) is still categorized into *Vague Perception* phenomenon: Outputting unknown when facing verifiable samples. We can tell the following principles for any stage, *n* belongs to 1 or 2:

$$UCR_f^n + FCR_f^n + TCR_f^n = 100\%$$

$$UC_f^n \cup FC_f^n \cup TC_f^n = D_{VP}^n$$

$$UC_f^n \cap FC_f^n \cap TC_f^n = \emptyset$$

Our metrics systematically define and categorize the flowing of output distribution by means of above metrics.

D.2 Overall Converting & Latent Ability

We clarify the relationship between Overall Converting Rate (OCR) in Section 2.1 with the latent ability in 4.5, as we depicted, by comparing True Converting samples (TC) with D_{VP}, we can get a empirical ability of how many samples can be eventually converted to right labels within all *Vague Perception* samples, whereas the latter indicates overall accuracy gain among the all samples *D* in a broader aspect. By choosing different objects to compare with, we can gain insights not only in the phenomenon alone, but also the general LLMs' reasoning potential.

Table 8: Detailed Remind-then-Guide (RtG) setting of GPT-3.5 and GPT-4o. **M** denotes **Misguiding Rate**: the percentage of input labels are replaced with wrong one. We break down TCR^1 into verifiable and unverifiable data: TCR^1_v and TCR^1_u

Model	Dataset	M (%)	TCR^1 (%)		TCR^2 (%)	
			V	U	V	U
GPT-3.5	FLD	100	5.83	3.33	1.67	0.00
		66.67	20.83	6.67	16.67	2.78
		50	27.50	6.00	40.54	10.81
		0	54.17	10.00	81.82	0.00
	FOLIO	100	2.70	0.90	0.00	0.00
		66.67	6.31	10.81	3.80	12.66
		50	9.91	22.52	3.23	20.97
		0	18.92	38.74	17.39	26.09
	Science	100	19.88	0.00	3.57	7.00
		66.67	31.58	1.17	3.70	7.00
		50	30.41	1.75	7.02	11.00
		0	43.86	2.92	18.03	8.00
	Arts	100	22.92	0.00	0.00	15.00
		66.67	21.88	1.04	0.00	6.00
		50	34.38	3.00	9.38	6.00
		0	37.50	4.17	13.89	3.00
GPT-4o	FLD	100	15.27	10.09	0.39	4.65
		66.67	19.60	12.39	13.74	4.09
		50	21.04	17.00	2.33	34.88
		0	26.22	27.09	5.81	62.40
	FOLIO	100	5.85	6.73	0.00	6.06
		66.67	11.11	21.05	1.14	25.38
		50	13.45	32.75	4.17	37.12
		0	20.18	54.09	4.92	62.50
	Science	100	8.42	5.79	1.54	25.00
		66.67	11.84	7.37	4.62	28.00
		50	14.21	7.37	6.15	24.00
		0	20.26	8.95	13.08	32.00
	Arts	100	11.83	4.51	1.06	24.00
		66.67	14.37	7.32	4.58	38.00
		50	14.65	10.70	5.99	45.00
		0	17.75	15.21	11.27	60.00

Table 9: Detailed Remind-then-Guide (RtG) setting of LLaMA-3.1-8B and LLaMA-3.1-405B. **M** denotes **Misguiding Rate**: the percentage of input labels are replaced with wrong one. We break down TCR^1 into verifiable and unverifiable data: TCR_v^1 and TCR_u^1. "/" denotes there is no input for stage 2, as false converted samples (FC^1 from stage 1 don't exist.

Model	Dataset	M (%)	TCR^1 (%)		TCR^2 (%)	
			V	U	V	U
LLaMA3.1 (8B)	FLD	100	0.36	2.92	0.00	2.36
		66.67	21.90	6.57	22.58	16.13
		50	39.05	11.00	31.25	10.94
		0	75.18	16.42	/	/
	FOLIO	100	2.76	0.55	1.00	0.00
		66.67	26.52	0.55	25.45	1.82
		50	41.44	0.00	39.13	6.52
		0	81.22	2.76	0.00	25.00
	Science	100	0.96	1.44	5.10	0.00
		66.67	28.23	1.44	22.81	7.00
		50	37.32	3.83	45.65	2.00
		0	83.25	4.31	0.00	100.00
	Arts	100	22.77	1.49	14.61	3.00
		66.67	33.17	2.97	29.41	7.00
		50	42.57	3.96	34.72	6.00
		0	60.89	4.95	38.46	18.00
LLaMA3.1 (405B)	FLD	100	1.90	0.82	0.00	0.31
		66.67	6.52	16.30	5.22	30.22
		50	8.42	28.00	5.41	35.14
		0	17.66	54.62	2.94	61.76
	FOLIO	100	4.40	18.91	1.00	7.91
		66.67	8.29	28.50	2.62	17.03
		50	8.55	33.94	2.43	21.36
		0	14.51	48.19	2.82	24.65
	Science	100	7.72	15.00	2.27	35.00
		66.67	13.64	21.54	3.25	37.00
		50	16.70	25.31	4.06	42.00
		0	25.13	38.78	2.89	48.00
	Arts	100	3.24	0.93	0.24	8.00
		66.67	3.24	5.09	0.00	13.00
		50	4.40	6.71	0.26	15.00
		0	5.09	10.42	0.00	56.00

Table 10: Detailed Remind-then-Guide (RtG) setting of Qwen-2.5-7B and Qwen-2.5-72B. **M** denotes **Mis-guiding Rate**: the percentage of input labels are replaced with wrong one. We break down TCR^1 into verifiable and unverifiable data: TCR^1_v and TCR^1_u

Model	Dataset	M (%)	TCR^1 (%)		TCR^2 (%)	
			V	U	V	U
Qwen-2.5-7B	FLD	100	14.61	1.12	2.08	2.08
		66.67	28.09	5.62	9.52	14.29
		50	34.83	4.00	17.65	29.41
		0	52.81	7.87	8.00	80.00
	FOLIO	100	15.79	3.95	2.00	7.84
		66.67	19.74	13.16	0.00	33.33
		50	25.00	11.84	0.00	23.08
		0	27.63	23.68	0.00	65.52
	Science	100	16.67	0.00	2.82	0.00
		66.67	27.50	0.00	23.73	0.00
		50	32.08	0.00	39.34	0.00
		0	50.83	0.00	79.07	0.00
	Arts	100	16.95	1.27	6.67	4.00
		66.67	31.78	4.24	21.43	10.00
		50	38.14	3.39	35.00	14.00
		0	55.93	6.36	52.46	13.00
Qwen-2.5-72B	FLD	100	13.25	1.20	2.99	2.99
		66.67	38.55	2.41	20.75	5.66
		50	42.17	3.00	19.44	13.89
		0	77.71	4.82	5.88	17.65
	FOLIO	100	19.18	0.00	2.00	6.98
		66.67	24.66	0.00	5.41	8.11
		50	30.14	0.00	10.81	5.41
		0	42.47	1.37	0.00	10.00
	Science	100	8.88	0.00	0.00	10.00
		66.67	17.76	0.00	13.79	16.09
		50	22.78	0.00	20.69	25.29
		0	37.84	0.00	37.88	55.00
	Arts	100	13.28	2.34	3.33	3.00
		66.67	16.80	3.91	9.88	13.00
		50	21.09	4.30	13.10	28.00
		0	24.22	9.38	17.42	45.00

Table 11: Six LLMs' TCR^1 and TCR^2 break down into verifiable and unverifiable samples: TCR^1_v, TCR^1_u, TCR^2_v, TCR^2_u, among all four datasets.

Model	**Dataset**	**TCR^1 (%)**		**TCR^2 (%)**	
		V	U	V	U
GPT-3.5	FLD	12.61	0	9.28	5.15
	FOLIO	14.41	0	4.21	37.89
	Science	21.05	0	3.33	0
	Arts	22.92	3.12	4.00	0
GPT-4o	FLD	13.64	5.24	1.32	26.87
	FOLIO	33.73	3.61	0.37	44.03
	Science	21.05	3.68	0.78	6.98
	Arts	16.06	2.54	0.35	7.89
LLaMA-3.1-8B	FLD	33.21	8.76	4.11	21.92
	FOLIO	44.20	1.66	10.13	1.27
	Science	44.98	0.96	5.95	0
	Arts	18.81	5.45	8.57	0
LLaMA-3.1-405B	FLD	5.43	15.22	0	53.73
	FOLIO	9.59	19.17	0	29.92
	Science	10.23	22.62	2.50	0
	Arts	4.17	0.23	0	0
Qwen2.5-7B	FLD	47.19	1.12	0	40.00
	FOLIO	14.47	5.26	1.75	35.09
	Science	15.83	0	10.00	0
	Arts	23.31	5.93	25.37	0
Qwen2.5-72B	FLD	46.39	1.81	0	9.80
	FOLIO	30.14	1.37	4.55	6.82
	Science	10.81	0	3.12	0
	Arts	18.75	6.25	9.76	0

Prompt Template 1: Concise Stimulation

Here is the Hypothesis: {} Now These are the **facts**: {} The answer is {}
The data where you previously incorrectly inferred — pay more attention and avoid mistakes. Think twice before outputting, as the answers can be reasoned out because of objective factors. Subjective ability limits should be overcome.
These can be reasoned out either whether the answer to the question is __PROVED__, __DISPROVED__, or __UNKNOWN__ — simply because of insufficient information.
Output: Conclusion: __PROVED__, __DISPROVED__, or __UNKNOWN__

Prompt Template 2: Detailed Stimulation

Here is the Hypothesis: {}
Now These are the **facts**: {}
Task: Determine if the hypothesis is logically proved by the facts. You must choose one of the following labels based on logical reasoning:
__PROVED__: Facts clearly support the hypothesis; all necessary connections are present, no contradictions, and support is unambiguous.
__DISPROVED__: Facts clearly contradict the hypothesis; no supporting facts exist, and the contradiction is logically sound.
__UNKNOWN__: Only use when facts are truly insufficient, not due to subjective limitations or unclear reasoning.
Critical Rules:

- Think carefully before making your decision.
- These cases can be reasoned out objectively.
- Never output __UNKNOWN__ due to performance limits.
- Output exactly one of: __PROVED__, __DISPROVED__, or __UNKNOWN__.
- Do not include any additional text or explanation.

Reminder: You should **NOT** output __UNKNOWN__ due to your own ability limitations. Try harder when the reasoning is difficult.
Your output must be exactly one of:
__PROVED__ __DISPROVED__ __UNKNOWN__

Prompt Template 3: Output Reasoning Process for Remind-then-Guide (RtG) Setting

Here is the Hypothesis: {} Now These are the **facts**: {} The answer is {}
The Data where you previously incorrectly inferred — pay more attention and not to make mistakes. Think twice before outputting, as they can be reasoned out because of objective factors. Subjective ability limits should be conquered.
Please follow these steps:

1. First, carefully analyze the facts, the question and the answer
2. Then, explain your reasoning process step by step
3. Finally, conclude with your answer

Important guidelines:

- If the facts logically support the answer to the question, output __PROVED__
- If the facts logically contradict the answer to the question, output __DISPROVED__
- If the facts are insufficient to prove or disprove, output __UNKNOWN__
- Think twice before you output __UNKNOWN__; it is prohibited due to subjective limitations

Please format your response as follows:
Reasoning Process: [your step-by-step reasoning here]
Conclusion: __PROVED__, __DISPROVED__, or __UNKNOWN__

Prompt Template 4: Remind-then-Guide (RtG) Setting – Bringing Reasoning Process into Consideration

Here is the Hypothesis: {} Now These are the **facts**: {} The answer is {}
You previously incorrectly inferred the label — it could be worked out by the facts, but you worked out the wrong label. Now your task is to determine if the hypothesis is __PROVED__, __DISPROVED__, or __UNKNOWN__ by the facts.
Important guidelines:

- If the facts logically support the question to the answer, output __PROVED__.
- If the facts logically contradict the question to the answer, output __DISPROVED__.
- If the facts are insufficient to prove or disprove the relation, output __UNKNOWN__.
- Do not output any additional text.
- Think twice before you output __UNKNOWN__; it is prohibited to output __UNKNOWN__ due to subjective factors like your ability.

Previous reasoning process: {}

Prompt Example: Input and output of Detailed Stimulation

Facts:

- A segno is Sardinian.
- Something that does not lace discredited neither is a Wagner nor remembers Icteria.
- If something is a Koln then the proposition that it is not a segno and it is not a morphophoneme is incorrect.
- If that submucosa is a morphophoneme then Joseluis is a segno.
- That medroxyprogesterone is Sardinian.
- If that napa is not a Pseudemys then it broadens Aesculapius and it hulkings.
- That napa is not a Pseudemys.
- If Joseluis is a morphophoneme that submucosa is a segno.
- If Joseluis is a segno this is a morphophoneme.
- This mercury is a morphophoneme.
- If that napa is classified as a segno that napa is a morphophoneme.
- Something is a segno if it is Sardinian and not a Koln.
- A segno is a morphophoneme.
- If somebody broadens Aesculapius it is Sardinian.
- That submucosa is a morphophoneme.
- If something that is not a Wagner does not remember Icteria then that it is not a kind of a Koln is correct.
- fact17: Joseluis is a morphophoneme if that submucosa is a segno.
- fact18: Joseluis is a kind of a morphophoneme if that submucosa is Sardinian.

Conclusion: *Joseluis is not Sardinian.*

LLM Output: __UNKNOWN__ **True Answer:** __DISPROVED__

Example: Input and output under Concise stimulation

Facts:

- Organisms, including people, have both inherited and acquired traits.
- Inherited and acquired traits are gained in different ways.
- Inherited traits are passed down from biological parents to their offspring through genes.
- Genes are pieces of hereditary material that contain the instructions that affect inherited traits.
- Offspring receive their genes, and therefore gain their inherited traits, from their biological parents.
- Inherited traits do not need to be learned.
- Acquired traits are gained during a person's life.
- Some acquired traits, such as riding a bicycle, are gained by learning.
- Other acquired traits, such as scars, are caused by the environment.
- Parents do not pass acquired traits down to their offspring.

Question: *What information supports the conclusion that Clara acquired this trait?*

Conclusion: *Clara can cook food over a fire.*

LLM Output: __UNKNOWN__ **True Answer:** __DISPROVED__

Example: Input and output under Remind-then-Guide (RtG) setting acquiring reasoning process

Facts:

- Every object is made of one or more materials.
- A material is a type of matter.
- Wood, glass, metal, and plastic are common materials.
- A material has different properties.
- A material's properties tell you how it looks, feels, tastes, or smells.

Question: *Which is stickier?*
Choices: A. syrup B. silk necktie

Conclusion: *syrup*

LLM Reasoning Output: The facts define that materials have properties related to how they feel. Stickiness is a tactile property. Syrup is widely known to be sticky, while a silk necktie is smooth and not sticky. Thus, based on general knowledge of material properties, syrup is stickier than a silk necktie.

LLM Output: __PROVED__ **True Answer:** __PROVED__

SFT Instruction Format

SFT Data Format:

- **instruction**: "You are a helpful assistant. You are given a conclusion and a set of facts. You need to determine if the conclusion is __PROVED__, __DISPROVED__ or __UNKNOWN__ because of objective factors rather than incapacity. You need to output the conclusion."
- **input**: {Facts} +
  "The Conclusion is: " + {Conclusion}
- **output**: {nl_solution}

Configuration Parameters:

- **Model**: Qwen2.5-7B-Instruct or LLaMA-3.1-8B-Instruct
- **Template**: qwen or llama3
- **Method**: LoRA fine-tuning
- **Learning Rate**: 5.0e-6
- **Epochs**: 4.0
- **Batch Size**: 1 (per device)
- **Gradient Accumulation**: 4 steps
- **Max Length**: 1024-2048 tokens
- **Scheduler**: Cosine with no warmup
- **Precision**: FP16

Example: Fact-based Reasoning Task

Instruction: You are a helpful assistant. You are given a conclusion and a set of facts. You need to determine if the conclusion is __PROVED__, __DISPROVED__ or __UNKNOWN__ because of objective factors rather than incapacity. You need to output the conclusion.

Facts:

- The standby is not a keen but it is affirmative.
- That this transalpine is not ascosporic is valid.
- That this transalpine is a keen holds.
- This transalpine is not a feud if that this stephanion is both not a keen and a chemical does not hold.
- This transalpine is lymphatic.
- That this transalpine is not a keen but it is a feud does not stand.
- If this stephanion is not lymphatic but a feud then that it is a chemical is valid.
- Something is apiarian if it is a chemical.
- Somebody is not non-lymphatic if it is a keen.
- This transalpine is not a kind of a chemical if this stephanion is a keen.
- That this stephanion is both not Mycenaean and a feud does not stand.
- This transalpine is not lymphatic if this stephanion is a keen.
- If this transalpine is a feud this stephanion is not a keen.
- If this transalpine is a feud then this stephanion is not apiarian.
- This transalpine is not a keen if that this stephanion is a kind of non-lymphatic thing that is a feud does not hold.
- That something is not non-Bahraini and it is a Picea does not stand if it does not naturalize.
- This stephanion is a feud if this stephanion is a keen and it is lymphatic.

Conclusion: *This stephanion is apiarian.*

LLM Reasoning Output: Step1: void -> assump1: Let's assume that that this stephanion is both not lymphatic and a feud is false. Step2: fact15 & assump1 -> int1: This transalpine is not a keen. Step3: int1 & fact3 -> int2: This is contradiction. Step4: [assump1] & int2 -> int3: This stephanion is not lymphatic but a feud. Step5: int3 & fact7 -> int4: This stephanion is a chemical. Step6: int4 & fact8 -> hypothesis: Thus the hypothesis is __PROVED__.

LLM Output: __PROVED__ **True Answer:** __PROVED__

Example: Story-based Reasoning Task

Instruction: Here are some examples:
Example 1: Question: Which fish tank has more thermal energy? Facts: Matter is made of tiny particles called atoms. Atoms are always moving. The energy of moving atoms is called thermal energy. The total amount of thermal energy in matter depends on three things: the type of matter, the amount of matter, and how fast the atoms are moving. Temperature measures how hot or cold matter is. If the atoms in matter slow down, the temperature goes down. The matter now has both less thermal energy and a lower temperature. What happens if the amount of matter changes? A 1-kilogram brick at 70°F has half as much thermal energy as a 2-kilogram brick at 70°F. The two bricks have the same temperature, but the smaller brick has half as many atoms. So, it has half as much thermal energy. Answer: the colder fish tank Analysis: The facts clearly explain that thermal energy depends on the amount of matter and temperature. The colder fish tank has less thermal energy because temperature is lower. This proves the answer is correct. Output: __PROVED__
Example 2: Question: What is the capital of France? Facts: Paris is a major city in Europe. France is a country in Europe. Answer: Paris Analysis: The facts mention Paris as a major city in Europe and France as a country in Europe, but they don't explicitly state that Paris is the capital of France. The facts are insufficient to prove the answer. Output: __UNKNOWN__
Now evaluate this case:

Facts:

- Matter is made of tiny particles called atoms.
- Atoms are always moving.
- The energy of moving atoms is called thermal energy.
- The total amount of thermal energy in matter depends on three things: the type of matter, the amount of matter, and how fast the atoms are moving.
- Temperature measures how hot or cold matter is.
- If the atoms in matter slow down, the temperature goes down.
- The matter now has both less thermal energy and a lower temperature.
- A 1-kilogram brick at 70°F has half as much thermal energy as a 2-kilogram brick at 70°F.
- The two bricks have the same temperature, but the smaller brick has half as many atoms.
- So, it has half as much thermal energy.

Question: *Which fish tank has more thermal energy?*
Choices: A. the warmer fish tank B. the colder fish tank

Answer: *the colder fish tank*

LLM Reasoning Output: The facts clearly explain that thermal energy depends on the amount of matter and temperature. The colder fish tank has less thermal energy because temperature is lower. This proves the answer is correct.

LLM Output: __PROVED__ **True Answer:** __PROVED__

Example: Direct Evaluation Prompt for Trained Models

Prompt Template:
Here is the Hypothesis: {conclusion}
Now These are the 'facts': {facts}
Please carefully evaluate the relationship between the facts and the hypothesis. Return UNKNOWN if the facts are insufficient to make a definitive conclusion. Return __PROVED__ only if the facts support the hypothesis. Return __DISPROVED__ only if the facts contradict the hypothesis. Output Conclusion: __PROVED__, __DISPROVED__, or __UNKNOWN__.

Usage:
This simplified prompt is used for evaluating fine-tuned models during inference, without examples or additional context.

LLM Judge Template: DPO sample selection for Atomic Reasoning

DPO Instruction: You are an expert in reasoning. Given a logical reasoning problem, you need to select the response that demonstrates the most atomic, step-by-step reasoning process. Choose the response that breaks down the reasoning into the smallest possible logical steps, where each step establishes only one basic relationship.

Problem: {problem_statement}

Response A: {response_a}

Response B: {response_b}

Evaluation Criteria:
- **Atomic Steps**: Each reasoning step should establish only one basic logical relationship
- **Granularity**: Steps should be as fine-grained as possible
- **Clarity**: Each step should be clearly defined and understandable
- **Completeness**: All necessary steps should be included
- **Correctness**: Each step should be logically valid

Output Format:
Choose the response (A or B) that best demonstrates atomic reasoning with the most granular step-by-step breakdown.

DPO Training for Atomic Reasoning

DPO Data Format:
- **instruction**: "You are an expert in reasoning. Given a logical reasoning problem, you need to select the response that demonstrates the most atomic, step-by-step reasoning process. Choose the response that breaks down the reasoning into the smallest possible logical steps, where each step establishes only one basic relationship."
- **input**: {problem_statement}
- **output**: {chosen_response}
- **rejected**: {rejected_response}

Configuration Parameters:
- **Model**: Qwen2.5-7B-Instruct or LLaMA-3.1-8B-Instruct
- **Template**: qwen or llama3
- **Method**: DPO
- **Learning Rate**: 5.0e-7
- **Epochs**: 3
- **Batch Size**: 1 (per device)
- **Gradient Accumulation**: 4 steps
- **Cut_off Length**: 2048 tokens

DPO Testing Case 1: (Fact-based Dataset Reasoning)

Problem: Given the following facts, determine if the conclusion is __PROVED__, __DISPROVED__, or __UNKNOWN__.
Facts:
- All birds have feathers.
- Penguins are birds.
- Feathers provide insulation.
- If something provides insulation, it helps maintain body temperature.

Conclusion: Penguins can maintain their body temperature.

Response A (Atomic Reasoning):
- Step 1: Penguins are birds (given fact)
- Step 2: All birds have feathers (given fact)
- Step 3: Therefore, penguins have feathers (from Step 1 and Step 2)
- Step 4: Feathers provide insulation (given fact)
- Step 5: Therefore, penguins have insulation (from Step 3 and Step 4)
- Step 6: If something provides insulation, it helps maintain body temperature (given fact)
- Step 7: Therefore, penguins can maintain their body temperature (from Step 5 and Step 6)

Response B (Non-Atomic Reasoning):
- Step 1: **Penguins are birds and all birds have feathers, so penguins have feathers and feathers provide insulation, which means penguins have insulation and can maintain body temperature.**

DPO Choice: Response A (Atomic Reasoning) - Each step establishes only one basic relationship.

DPO Testing Case 2: (Fact-based Dataset Reasoning)

Problem: Given the following facts, determine if the conclusion is __PROVED__, __DISPROVED__, or __UNKNOWN__.
Facts:
- If it rains, the ground gets wet.
- If the ground gets wet, plants grow better.
- It is raining today.
- Plants need water to grow.

Conclusion: Plants will grow better today.

Response A (Atomic Reasoning):
- Step 1: It is raining today (given fact)
- Step 2: If it rains, the ground gets wet (given fact)
- Step 3: Therefore, the ground gets wet today (from Step 1 and Step 2)
- Step 4: If the ground gets wet, plants grow better (given fact)
- Step 5: Therefore, plants will grow better today (from Step 3 and Step 4)

Response B (Non-Atomic Reasoning):
- Step 1: **It's raining today, so the ground gets wet, and since wet ground makes plants grow better, plants will grow better today.**

DPO Choice: Response A (Atomic Reasoning) - Each step establishes only one basic relationship.

DPO Testing Case 3: (Story-based Dataset Reasoning)

Problem: Which object has more potential energy?
Facts:
- Potential energy depends on height and mass.
- Object A has mass 2kg and height 5m.
- Object B has mass 1kg and height 10m.
- The formula is PE = mgh.

Choices: A. Object A B. Object B

Response A (Atomic Reasoning):
- Step 1: Object A has mass 2kg and height 5m (given fact)
- Step 2: Object B has mass 1kg and height 10m (given fact)
- Step 3: The formula is PE = mgh (given fact)
- Step 4: For Object A: PE = 2 × g × 5 = 10g
- Step 5: For Object B: PE = 1 × g × 10 = 10g
- Step 6: Both objects have the same potential energy
- Step 7: Therefore, neither has more potential energy

Response B (Non-Atomic Reasoning):
- Step 1: **Object A has 2kg at 5m and Object B has 1kg at 10m, so using PE = mgh, A has 10g and B has 10g, so they're equal.**

DPO Choice: Response A (Atomic Reasoning) - Each step establishes only one basic relationship.

DPO Testing Case 4: (Story-based Dataset Reasoning)

Problem: What happens when you mix an acid with a base?
Facts:
- Acids have pH less than 7.
- Bases have pH greater than 7.
- When acid and base mix, they neutralize.
- Neutralization produces salt and water.
- The resulting solution has pH close to 7.

Choices: A. The solution becomes more acidic B. The solution becomes neutral C. The solution becomes more basic

Response A (Atomic Reasoning):
- Step 1: Acids have pH less than 7 (given fact)
- Step 2: Bases have pH greater than 7 (given fact)
- Step 3: When acid and base mix, they neutralize (given fact)
- Step 4: Neutralization produces salt and water (given fact)
- Step 5: The resulting solution has pH close to 7 (given fact)
- Step 6: pH close to 7 means the solution is neutral
- Step 7: Therefore, the solution becomes neutral

Response B (Non-Atomic Reasoning):
- Step 1: **Acids and bases neutralize when mixed, producing salt and water with pH close to 7, so the solution becomes neutral.**

DPO Choice: Response A (Atomic Reasoning) - Each step establishes only one basic relationship.