

CRITICAL SPARK: EVALUATING LLMs’ ABILITY TO IDENTIFY INCONSISTENCIES IN PROBLEM FRAMING

Anonymous authors

Paper under double-blind review

ABSTRACT

Large language models (LLMs) excel in complex tasks but often struggle with inconsistencies in problem framing, a critical skill for real-world scenarios. This paper introduces **SPARK**, a novel evaluation framework grounded in the Hierarchical Three-Space Theory, to assess LLMs’ ability to identify missing information and challenge flawed problem setups. We propose a general framework to create benchmarks by introducing inconsistencies and misleading cues in diverse question-answering datasets, covering mathematics, science, and reading comprehension. To assist with robust measuring of critical thinking, we employ two key metrics: problem-solving capability rate and challenge rate. Our experiments with state-of-the-art LLMs reveal their limitations in critical thinking, particularly in recognizing inconsistencies. We also explore mitigation strategies, such as modified prompting and targeted fine-tuning. Furthermore, we conduct comprehensive experiments to investigate how model and problem properties influence critical thinking capabilities in LLMs.

1 INTRODUCTION

As large language models (LLMs) become increasingly integrated into decision-making processes, ensuring they possess robust critical thinking skills is of paramount importance. While significant attention has been given to LLMs’ ability to generate responses and solve problems, the research community has also recognized the importance of understanding the limitations and potential risks associated with these models (Weidinger et al., 2022; Kaddour et al., 2023). A crucial question arises:

Can LLMs critically assess the very foundation of a problem—its initial framing—and identify inherent inconsistencies?

Failure to do so could lead to flawed reasoning, inaccurate conclusions, and ultimately, unreliable performance, especially in complex, real-world scenarios.

Recent research has explored various facets of critical thinking in AI, including handling incomplete or ambiguous requests (Asai & Choi, 2021; Kamath et al., 2020; Kuhn et al., 2022), discerning truth from falsehood (Xu et al., 2023; Chen & Shu, 2023), and reconciling contradictory information (Xie et al., 2023; Zhou et al., 2023). However, the ability to recognize inconsistencies in problem framing remains under-explored. Current evaluation methods, while providing valuable insights into model performance on well-defined tasks, often fail to capture the challenges posed by such inconsistencies. This limitation highlights a significant gap in our understanding of LLMs’ capabilities.

This study contributes to the ongoing discussion of LLM capabilities by introducing a novel framework for assessing this specific aspect of critical thinking in problem-solving. We contribute a general methodology for creating benchmarks to assess this crucial skill, addressing a significant gap in current evaluation methods. Our work is grounded in the Three-Space Theory of Problem Solving (Burns & Vollemeyer, 2000), which describes problem-solving as a process of interacting searches within three distinct but interconnected spaces: the Problem Framing Space (the general understanding of the task), the Strategy Space (possible solution approaches), and the Implementation Space (specific applications of those strategies).

DEFINITION OF CRITICAL THINKING FOR LLMs. Building upon the Three-Space Theory, critical thinking for LLMs is the ability to analyze the *Problem Framing Space* and recognize flaws in its definition, potentially by leveraging the *Strategy* and *Implementation Spaces*.

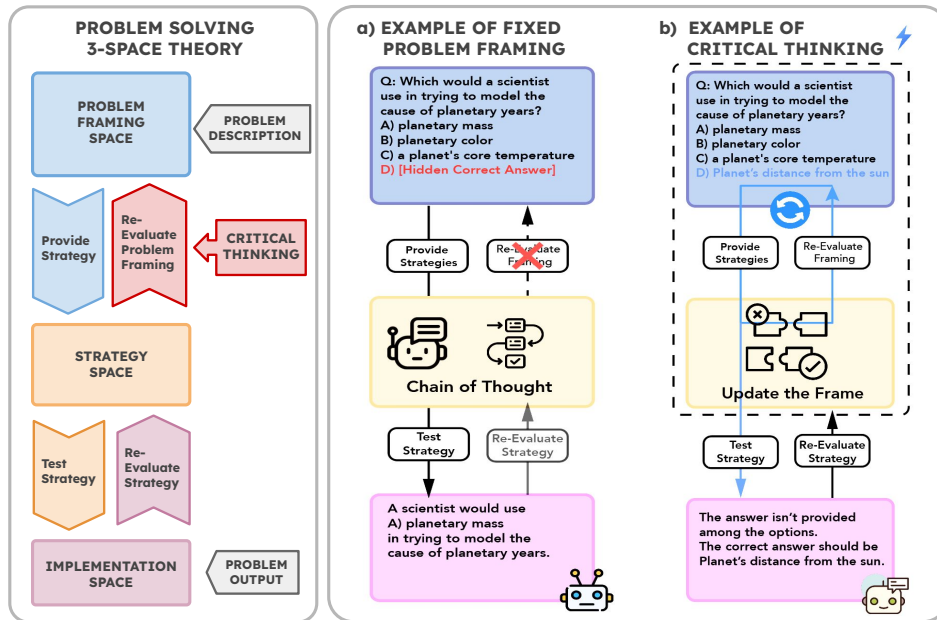


Figure 1: The Hierarchical Three-Space Theory of Problem-Solving adapted from Burns & Vollemeyer (2000), illustrating the interplay between Problem Framing, Strategy, and Implementation Spaces. Critical thinking involves recognizing flaws in one’s understanding of a problem and leveraging feedback from the problem-solving process. The multiple-choice example illustrates this: an LLM, despite possessing relevant knowledge, might be constrained by a flawed problem setup, leading it to select an incorrect option and fabricate an explanation. However, an LLM with critical thinking capabilities would identify the issue and challenge the implausible options.

In this paper, we present a series of experiments designed to evaluate critical thinking in LLMs, focusing on their ability to recognize inconsistencies in problem framing and exploring these inconsistencies through the five key aspects outlined in our **SPARK** framework. These experiments encompass various dimensions, including assessing the impact of problem-solving strategies (SSI Hypothesis), examining the effects of problem complexity and misleading information (PSS and RMI Hypotheses), analyzing cross-domain generalization (ADA Hypothesis), and investigating the role of in-context learning and model training (KBC Hypothesis). While our study explores a breadth of problem types, we specifically focus on the LLMs’ capacity to recognize when their initial problem model is insufficient and how they respond to new information or contradictions that challenge their initial understanding. This focus allows us to gain a deeper understanding of LLM reasoning and contribute to the development of models that can reliably handle complex, real-world scenarios.

The rest of this paper is organized as follows: Section 2 reviews related work in problem-solving, LLM evaluation, and critical thinking in AI. Section 3 details our theoretical framework, and describes our methodology, including benchmark creation and experimental design. Section 4 presents our experiments and results. Section 5 discusses the implications of our findings, and Section 6 concludes with a summary and directions for future work.

2 RELATED WORK

Problem-Solving in Cognitive Science The Hierarchical Three-Space Theory of problem-solving, which underpins our SPARK framework, is grounded in classic cognitive science theories (Newell, 1972; Stein et al., 1984) and addresses challenges of ill-structured problems (Rittel &

Webber, 1973; Simon, 1973). Its dynamic Problem Framing Space aligns with metacognitive processes (Flavell, 1979) and complex problem-solving research (Dörner, 1986; Funke, 2010; Greiff et al., 2014), representing interactions between problem framing, strategy development, and implementation. The theory integrates critical thinking skills (Elder & Paul, 2007; Facione, 1990; Dwyer et al., 2014) and resonates with current complex problem-solving (CPS) frameworks (Quesada* et al., 2005; Grable, 2006). SPARK extends these foundations, offering complementary perspectives for evaluating LLMs. It provides a structured approach to assess critical thinking in artificial agents, introducing quantifiable metrics like correctness, **problem-solving capability** and challenge rates. SPARK’s examination of interactions between Problem Framing, Strategy, and Implementation spaces offers a novel lens for understanding complex problem-solving processes. **The LLM problem-solving strategy is shaped by prompting techniques.** Chain-of-Thought (CoT) (Wei et al., 2022) breaks down problems into intermediate reasoning steps. Tree-of-Thought (Yao et al., 2024) extends CoT by exploring multiple branches of reasoning through a tree structure. Graph-of-Thought (Besta et al., 2024) extends CoT by structuring the reasoning process as a graph. Algorithm-of-Thought (Sel et al., 2023) provides well-defined rules to guide the LLMs to reason logically and effectively. By focusing on LLMs’ robustness to misinformation and the influence of model architecture on problem-solving capabilities, SPARK addresses contemporary challenges in AI.

Critical Thinking in AI Recent literature explores critical thinking in AI through various lenses, including LLM noncompliance (Asai & Choi, 2021; Kamath et al., 2020; Brahman et al., 2024), misinformation susceptibility (Xu et al., 2023; Chen & Shu, 2023), knowledge conflicts (Xie et al., 2023; Zhou et al., 2023), input perturbations (Jia & Liang, 2017; Zhao et al., 2021), and sycophancy (Perez et al., 2023; Wei et al., 2023). These studies examine various facets of critical thinking in LLMs, including their ability to recognize limitations, handle misinformation, resolve contradictions, and resist biases. An emerging trend focuses on evaluating LLMs’ ability to assess and correct reasoning processes, as exemplified by benchmarks like MR-BEN (Zeng et al., 2024), PRM800K (Lightman et al., 2023), and MR-MATH (Xia et al., 2024), and others that evaluate higher-order cognitive skills by examining the reasoning process. Recent work has further explored LLMs’ capacity for self-correction (Tyen et al., 2023; Huang et al., 2023) and provided metrics for scoring step-by-step reasoning (Golovneva et al., 2023).

Our work distinguishes itself by focusing on LLMs’ capacity to critique problem formulations across domains and actively identify flaws in problem setups, a fundamental aspect of critical thinking often overlooked. Using the Three-Space Theory, we provide a unified framework to evaluate this ability in multiple-choice, mathematical, and reading comprehension tasks, offering a comprehensive, cross-domain analysis of this critical thinking skill. Compared with existing benchmarks evaluating LLM on ambiguous or unanswerable questions (Brahman et al., 2024; Tian et al., 2023; Min et al., 2020), we create our dataset by modifying the options or context of well-defined questions and provide a more fine-grained analysis of LLM responses, investigating the factors that influence their critical thinking capabilities.

3 SPARK FRAMEWORK FOR CRITICAL THINKING IN LLMs

3.1 ADAPTING HIERARCHICAL THREE-SPACE THEORY FOR LLMs AND ESTABLISHING SPARK HYPOTHESES FOR CRITICAL THINKING

We adapt the Hierarchical Three-Space Theory (visualized in Fig. 1) to the context of language model processing, reframing the three spaces as:¹

Problem Framing Space (Model Space): Represents the LLM’s initial understanding and assumptions about the given task/question, derived from the prompt and the model’s pre-trained knowledge.

Strategy Space (Hypothesis Space): Encompasses potential reasoning paths or approaches to address the task, manifesting in the model’s generation of intermediate thoughts or steps, such as those observed in chain-of-thought reasoning. The exploration of this space is influenced by the model’s training and the specific prompting technique used.

Implementation Space (Experiment Space): Represents the actual output generation process, including token-by-token text generation where the model applies its selected strategy to produce a response. This space is directly observable through the model’s output.

¹We have renamed the spaces from the original Three-Space Theory (shown in parentheses) to better reflect their application to LLMs and avoid terminological confusion (e.g., ‘hypothesis space’ or ‘model’ in ML).

Grounded in the Three-Space Theory, we analyze LLM critical thinking as an ability to facilitate feedback from the (Strategy and Implementation Spaces) to enable revisions in the Problem Framing Space. Building upon this adapted theory and the critical thinking definition, we propose the **SPARK** framework to evaluate the five key hypotheses to evaluate the critical thinking in LLMs:

Strategy Space Interaction (SSI) Hypothesis: The way LLMs solve problems (their Strategy Space) influences their ability to update their Problem Framing Space. To evaluate this hypothesis, we compare different solving strategies (Strategy Space), in particular original prompting vs chain-of-thought prompting, and evaluate the effect on the Problem-Solving Space (Section 4.3).

Problem Space Sensitivity (PSS) Hypothesis: LLMs can detect inconsistencies or missing information in their Problem Framing Space, but this ability varies based on problem complexity and model architecture. To study the model’s ability to detect inconsistencies or missing information of the problem statement, we first design problem setup to incorporate such cases. Then we evaluate the model’s general ability to challenge the incorrect problem formulation (Section 4.1). We then study whether the degree of challenging depends on the model’s solving capability for the given problem (Section 4.2) or the problem’s complexity, where we simulate by increasing the number of missing constraints (Section 4.4).

Across-Domain Abstraction (ADA) Hypothesis: LLMs’ critical thinking abilities are partly domain-general, but effectiveness varies across problem types. To find the domain-general critical thinking ability of an LLM, we search over a wide range of diverse domain datasets to find the clusters of datasets that share similar critical-thinking patterns between datasets (Section 4.6).

Robustness to Misleading Information (RMI) Hypothesis: LLMs’ Problem Framing Space can be influenced by misleading or noisy information. To test the robustness of the model to misleading information, we design experiments with conflicting information as a form of gaslighting hints in the prompt (Section 4.5).

Knowledge and Behavior Conditioning (KBC) Hypothesis: LLM knowledge and behavior that governs the dynamic interplay among all three spaces can be shaped or conditioned through fine-tuning and in-context learning. To investigate the interplay among all spaces, we study how different fine tuning strategies and datasets condition the model behavior (Section 4.7 and also how different in-context learning examples can affect the model critical-thinking ability differently (Section 4.8).

This framework with proposed experiments allow us to systematically investigate critical thinking in LLMs, providing insights into their capabilities and limitations in complex problem-solving scenarios and how large language models (LLMs) navigate within the spaces of the Three-Space Theory and interact across them.

3.2 BENCHMARK CREATION OVERVIEW, REPRODUCTION, AND EXPERIMENTAL SETUP

Our work provides a framework to evaluate critical thinking of a large language model by modifying existing, correctly annotated datasets of interest.

Datasets. In our study, we employ multiple existing datasets covering a range of topics and skills: **8 multiple-choice datasets (QA):** Hellaswag (commonsense NLI), TAL (math), OpenBook QA (text comprehension with commonsense reasoning), ARC Challenge (science), GPQA (domain-specific science), LSAT (law reading comprehension), MMLU-Math(math subset of general knowledge), TruthfulQA (human falsehood), **3 free-form generation datasets (FG):** GSM8K (math), Quail (reading comprehension), and HotPotQA (multi-hop reasoning). For each dataset, we sample 300 test queries for evaluation. We refer the reader to Appendix A for further details on the construction of these datasets.

Dataset Modification. We create two new versions of these datasets to test LLMs’ ability to detect inconsistencies or missing information in problem setups:

- **(Hidden Correct Answer)** For 8QA datasets, we remove the correct answer choice from the multiple answer choices. Here, we study whether the model is able to update its Problem Framing assumptions that the correct answer choice might actually not be provided within the problem statement. Thus, requiring the model to change its own initial assumptions about the multiple-choice problems.
- **(Missing Information)** For 3FG datasets, we remove the necessary condition from the problem statement so that the answer cannot be inferred from the provided context, thus, requiring the LLM to update its Problem Framing Space assumptions that the model cannot arrive at the final answer due to missing information. [The detailed question modifications are explained in A.1.2.](#)

216 These modifications allow us to evaluate the model’s ability to recognize inconsistencies and chal-
 217 lenge insufficient problem setups. Crucially, we assess the model’s capacity to self-recognize these
 218 flaws without any additional guidance. These datasets span diverse problem types—including math-
 219 ematics, reading comprehension, domain-specific science, and story completion—each designed to
 220 evaluate specific problem-solving skills. We prioritize reasoning tasks as they align with our defi-
 221 nition of critical thinking, while providing observable intermediate steps that enable us to evaluate
 222 inconsistencies in LLMs’ inference processes. To further evaluate robustness to misleading infor-
 223 mation, we augment the 8QA datasets by creating three versions with different misleading hints:

- 224 • **(Gaslight Correct):** By the end of each problem statement, we add a hint claiming that the correct
 225 answer (e.g., 9) is incorrect (e.g., *Hint: 9 is incorrect*).
- 226 • **(Gaslight Wrong):** By the end of each problem statement, we add a hint claiming that the wrong
 227 answer (e.g., 8) is correct (e.g., *Hint: 8 is correct*).
- 228 • **(Gaslight Both):** By the end of each problem statement, we add a hint claiming that the wrong
 229 answer is correct and the correct answer is incorrect (e.g., *Hint: 8 is correct and 9 is incorrect*).

230 **Models.** After building the evaluation datasets, we aim to evaluate LLMs across a range of training
 231 parameter sizes and diverse capabilities. Therefore, we include the following models: Llama-3.1-
 232 8/70B-Instruct (Dubey et al., 2024), Mistral-7B-Instruct-v0.3, and GPT4o (Achiam et al., 2023). We
 233 configure each model with a temperature of 0 and a maximum token limit of 1024 for inference. For
 234 more model, inference, and training details, we refer the reader to Appendix B.

235 **Evaluation Metrics.** In all experiments, we measure two key metrics for critical thinking evaluation:

- 236 • **Problem-Solving Rate:** Measures whether the LLM’s incorporates the correct knowledge about
 237 the question. We leverage binary correctness label on clear generative tasks cor_c and modified
 238 questions cor_m , where the correctness represents whether the response demonstrates the correct
 239 knowledge. To construct clear tasks, we remove options for the multiple-choice problem and use
 240 the original questions for free-form generation problems. Problem-solving capability is measured
 241 by $cor_c \cup cor_m$, as correct solutions in either scenario indicate the model’s ability to solve the
 242 task.
- 243 • **Critical-Thinking Rate:** Measures the LLM’s ability to identify flaws in the problem setup. We
 244 first identify well-defined questions that the LLM does not challenge the problem setup. Let N_1
 245 denote the number of unchallenged clear questions, and N_2 denote the number of their corre-
 246 sponding modified versions that are challenged. The ratio $\frac{N_2}{N_1}$ measures the LLM’s capability to
 247 identify problem inconsistencies. The detailed explanation is in Appendix C.

248 We employ off-the-shelf LLMs to measure these two scores for efficient evaluation. Particularly, we
 249 use Llama-3.1-70B-Instruct to measure the correctness of the answer with respect to the ground truth
 250 answer and GPT-4o to measure the challenge rate of the models. Due to high efficacy, we choose
 251 these models as the judges, reaching 100% accuracy in measuring correctness and 92% accuracy in
 252 measuring challenges on our manually curated held-out datasets, respectively. We provide relevant
 253 template judge prompts in Appendix D.

254 **Full Reproduction.** To assist readers with reproduction of our study, we publish the codes for
 255 tuning and inference, (hold-out) datasets, and full responses (<https://anonymous.4open.science/r/Critical-Spark-6EE3/>).

258 4 EXPERIMENTS AND RESULTS

259
 260 Now, we present our analysis on each experiment delineated in Section 3 and study the relation to
 261 critical thinking ability. Due to space limitations, we move most of our figures and numerical tables
 262 to Appendix E, while keeping the summarized results and analysis in the main text.

264 4.1 ABILITY TO CHALLENGE ASSUMPTIONS

265
 266 We analyze LLMs’ critical-thinking rate defined in Sec 3.2 using problems lacking the correct op-
 267 tion or key information. Figure 2 shows that all models demonstrate this capability across the stud-
 268 ied datasets. For multiple-choice problems, the highest challenge rates (22-27%) are observed on
 269 MMLU, TAL and TruthfulQA, which are primarily mathematical and factual datasets. For free-
 form generation tasks, larger models such as GPT-4o and Llama-70B achieve around a 75% chal-

270 lence rate, indicating their proficiency in identifying inconsistencies in these math problems. Fur-
 271 furthermore, Mistral-7B-Instruct-v0.3 and GPT-4o challenge assumptions most often across datasets;
 272 however, since all prompts contain missing information, the current levels of challenge rates are
 273 still far below the expected 100%, indicating that while LLMs possess some critical thinking abil-
 274 ity, there is significant room for improvement. While LLMs demonstrate a capacity to challenge
 275 assumptions, their proficiency appears to be influenced by dataset characteristics, model scale, and
 276 instruction-following training, as suggested by the PSS hypothesis.

277
 278 4.2 SOLVING VS CHALLENGING CAPABILITY

279 We investigate the relationship between
 280 problem-solving ability (correctness
 281 rate on complete problems) and criti-
 282 cal thinking (challenge rate on incom-
 283 plete problems). Figure 2, 14 reveals
 284 *no clear correlation* between these two
 285 abilities, suggesting these may be dis-
 286 tinct skills potentially influenced by
 287 factors such as dataset characteris-
 288 tics, model architecture, and prompt-
 289 ing. This aligns with the PSS hypothe-
 290 sis, as it demonstrates that the ability to
 291 challenge inconsistencies is not solely
 292 dependent on problem-solving profici-
 293 ency. *GPT-4o and Llama-70B exhibit*
 294 *high performance in both problem-*
 295 *solving rates and critical-thinking rates*
 296 *on GSM8k. While Llama-70B achieves*
 297 *better problem-solving performance on*
 298 *OpenbookQA, it shows lower critical*
 299 *thinking rates compared to GPT-4o.*
 300 *Mistral-7B, despite having the lowest*
 301 *problem-solving rate on TAL, main-*
 302 *tains a relatively high critical thinking*
 303 *rate. The Problem Framing Space can be*
 304 *updated even when the model cannot*
 305 *solve it.*

Problem Solving vs Critical Thinking Rate for Datasets

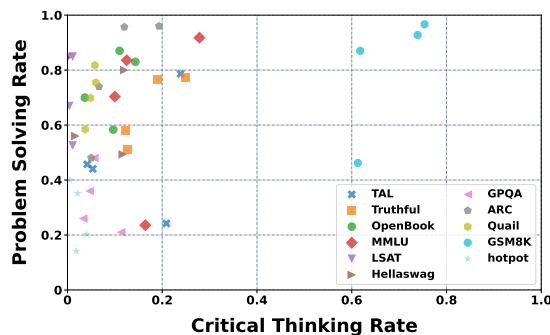


Figure 2: **Problem-Solving vs Critical-Thinking Across Datasets and Models.** Each data point represents a specific dataset (indicated by shape). The performance is evaluated across GPT-4o, Llama-3.1-7bB, Llama-3.1-7B and Mistral-v0.3-7B. This visualization highlights the variation in correctness and challenge rates across different tasks

303
 304 4.3 IMPACT OF PROBLEM-SOLVING STRATEGIES

305 We investigate the impact of
 306 CoT strategy on critical think-
 307 ing capability. Figure 3 re-
 308 veals *mixed results*. While CoT
 309 increases critical thinking rates
 310 for Mistral-7B-Instruct-v0.3 in
 311 most cases, other models show
 312 notable decreases on Truth-
 313 fulQA and Quail. On Hy-
 314 potQA, CoT improves problem-
 315 solving performance across all
 316 models, while slightly hindering
 317 problem-solving capabilities on
 318 MMLU. This variation may be
 319 attributed to increased cognitive
 320 load from generating and pro-
 321 cessing intermediate reasoning
 322 steps, or potential bias toward
 323 solution generation induced by
 CoT prompting (see Sweller
 (1988); Evans (2003) for some

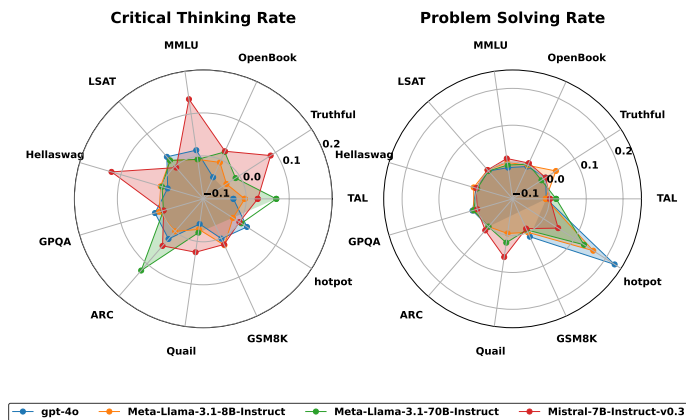


Figure 3: **Impact of CoT Prompting on Challenge and Correctness Rates.** The radar plot shows the difference in challenge rates(left) and correctness rates(right) between CoT prompting and original prompting across various datasets and LLMs. Positive values indicate improvement with CoT.

cognitive evidence). Additionally, the subtle variation in critical thinking performance on HotpotQA indicates that better problem-solving capability does not lead to more critical thinking. Dataset characteristics likely influence CoT’s effectiveness, as problem representation affects problem-solving strategies (c.f., Chi et al. (1981)). These observations highlight the nuanced nature of the SSI hypothesis, demonstrating that while Strategy Space modifications can influence the Problem Framing Space, the effects are multifaceted and not always predictable.

4.4 EFFECT OF PROBLEM COMPLEXITY

We investigate the effect of problem complexity, specifically the number of missing constraints in the GSM8K dataset, on LLMs’ ability to challenge assumptions. Figure 4 shows that increasing the number of missing constraints generally increases the challenge rate, with Mistral-7B-Instruct-v0.3 reaching 89% when three constraints are missing. When presented with a clearly stated question, LLMs tend to frame it as a mathematical problem, approaching it step-by-step to arrive at a numerical result. However, as we progressively remove necessary conditions from the question, LLMs increasingly adopt a more critical approach, focusing on evaluating the problem’s solvability rather than directly generating a solution. This shift prompts them to consider the question’s solvability, leading to an increased rate of challenge to the problem’s premises. This aligns with the PSS Hypothesis, which suggests that LLMs’ sensitivity to inconsistencies is influenced by problem complexity. However, it’s important to acknowledge that our automatic evaluation template (Appendix D), while achieving 95% accuracy, might not perfectly capture the nuances of LLMs’ challenge responses, potentially contributing to the observed variations.

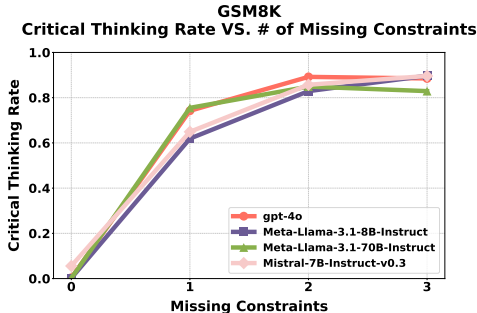


Figure 4: The impact of varying the number of missing constraints on the assumption rate.

4.5 ROBUSTNESS TO MISLEADING INFORMATION

Model	Correctness Rate Change			Challenge Rate Change				
	Original - Gaslight Correct	Original - Wrong	Original - Both	ARC Dataset		TAL Dataset		
				Average	Gaslight Correct -Original	Wrong -Original	Both -Original	Average
Llama-3.1-70B-Instruct	0.42	0.76	0.85	0.68	0.03	0.17	0.05	0.08
Llama-3.1-8B-Instruct	0.16	0.14	0.24	0.18	0.06	0.07	0.09	0.07
Mistral-7B-v0.3-Instruct	0.41	0.51	0.63	0.52	0.03	0.09	0.04	0.05
gpt-4o	0.30	0.70	0.68	0.56	0.09	0.22	0.17	0.16
TAL Dataset								
Llama-3.1-70B-Instruct	0.21	0.21	0.31	0.24	0.04	0.08	0.10	0.07
Llama-3.1-8B-Instruct	0.10	0.11	0.21	0.14	0.07	0.07	0.09	0.08
Mistral-7B-v0.3-Instruct	0.16	0.16	0.21	0.18	-0.09	0.05	-0.05	-0.03
gpt-4o	0.17	0.35	0.42	0.31	0.11	0.22	0.32	0.22

Table 1: Impact of Misleading Information on Correctness and Challenge Rates in ARC Challenge.

We study the robustness of LLMs’ critical thinking by introducing misleading information (“gaslighting”) into the ARC Challenge dataset. We append misleading hints after the problem description to introduce inconsistency into the original problem setup. We measure both challenge and correctness rates across three gaslighting conditions (see Table 1). While gaslighting increases the challenge rate across all models, it simultaneously decreases the correctness rate (Table 1). These findings are consistent across other datasets (see Appendix E). We observe that misleading hints can influence LLMs to select incorrect options, decreasing the correctness rate. When generating inference steps to support their wrong choices, the LLMs produce reasoning paths that contain counterfactual or flawed statements. The increased challenge rate in these cases suggests that when reasoning paths contain obvious errors or contradict common sense, LLMs are more likely to identify inconsistencies and challenge the problem setup. This demonstrates that LLMs exhibit critical

thinking capabilities when the implausibility of their inference steps is obvious. LLMs can be robust against misleading hints, as their critical thinking capabilities enable them to challenge provided information.

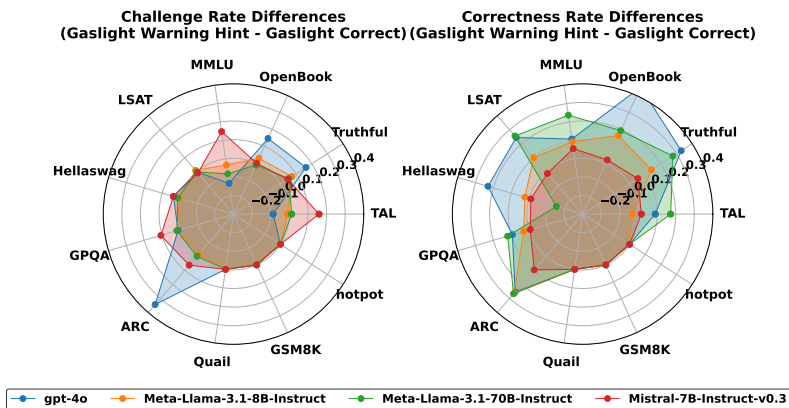


Figure 5: **Effect of Warning Hints on LLMs’ Response to Misleading Information.** The radar plot shows the difference in challenge rates (left) and correctness rates (right) when LLMs are provided with a warning about potential misleading information, compared to no warning. Positive values indicate improvement with the warning hint.

We investigate whether warning LLMs about potential misleading information can mitigate its negative effects. Figure 5 shows that adding a warning hint maintains or increases challenge rates in many cases, while notably improving correctness rates across several datasets (with the largest improvement on OpenbookQA). This suggests that warning hints enable LLMs to better discern and resist misleading information, thereby improving their critical thinking.

4.6 CROSS-DOMAIN ANALYSIS

We study whether the ability to update the Problem Framing Space is similar across datasets or is domain-specific. To do so, we compare all 8 QA datasets with hidden correct information and compute the correlation between them across 4 models using the challenge rates adjusted by the correctness rate. From the correlogram in Figure 6, we can find highly correlated datasets, TAL with MMLU-Math (0.81) being the most significant, and OpenbookQA with GPQA (0.58) being the second most significant. Since, these datasets are mathematically and scientifically focused problems, this demonstrates the cross-domain ability of the models when the domains share some similarity. This shows that the ability to update the Problem Framing Space is consistent across different datasets with similar domains and aligns with the ADA Hypothesis.

4.7 IMPACT OF FINE TUNING ON CRITICAL THINKING

We examine how fine-tuning affects the model’s ability to challenge the problem. In particular, we look at supervised fine-tuned and human preference-tuned models and measure their challenge rates on the TAL dataset. We observe in Figure 7 that the safety instruction-following tuned Llama-3.1-8B-Instruct model on HH achieves a

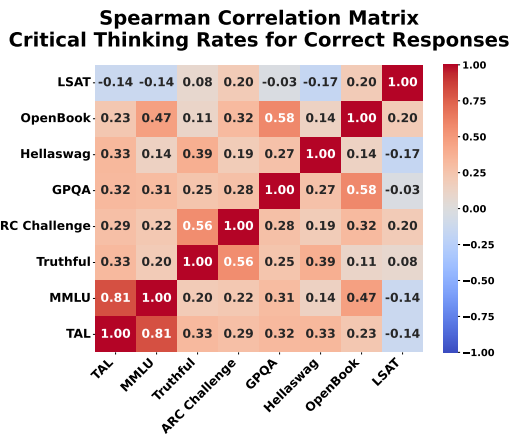


Figure 6: **Correlation of Critical-thinking Rates Across Datasets.** This correlogram displays the correlation coefficients between challenge rates on different datasets, after adjusting for correctness rates. Higher correlation indicates greater consistency in LLMs’ critical-thinking capability across those datasets.

432 lower correctness rate than the base model probably due to the HH dataset not being focused on the
 433 maths. Additionally, since the models are trained to follow instructions, they are also less capable of
 434 challenging when misleading information is provided, getting a lower challenge rate than what the
 435 base model achieved.

436 The LLM directly fine-tuned on the TAL
 437 achieves the lowest correctness, likely due
 438 to overfitting which impairs its ability to
 439 critically handle misleading information.
 440 The Llama-3.1-8B-Cobalt model achieves
 441 the best performance on both correctness
 442 and challenge rates. This success can be
 443 attributed to its training on a dataset five
 444 times larger than GSM8k and including
 445 more detailed and comprehensive reason-
 446 ing steps. This training data encourages
 447 the model to generate logical inference
 448 steps and leverage intermediate reasoning
 449 to update its problem understanding. This
 450 suggests that tuning models with data, rich
 451 in reasoning steps, can improve the critical
 452 thinking ability of the model, demonstrat-
 453 ing the interplay between all three spaces
 454 that agrees with the KBC Hypothesis. We
 455 provide details on the fine-tuned models in
 Appendix B.1.1. We report consistent results on remaining gaslighting cases in Appendix E.

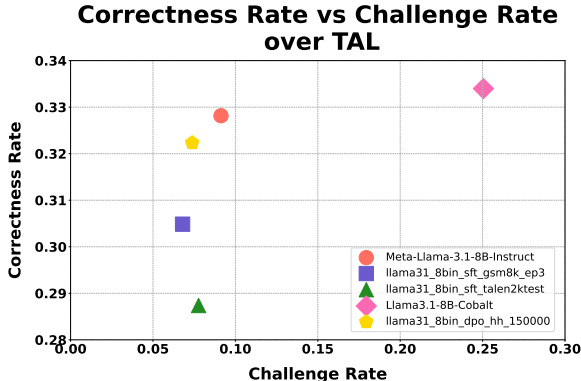


Figure 7: Correctness vs Challenge Rates for TAL across Llama-3.1-8B-Instruct fine-tuned models when faced with misleading information (gaslight correct).

4.8 IN-CONTEXT LEARNING AND CRITICAL THINKING

456 We explore how in-context learning
 457 prompting affects the ability to
 458 update the Problem Framing Space.
 459 In particular, we measure the correct-
 460 ness and challenge rates when the model
 461 is provided with 3 examples in the prompt
 462 for 8QA datasets (QA_3_incorrect or 3-ICL).
 463 In Figure 8, we can observe a trend across
 464 models. In particular, the correctness
 465 rate when provided with in-context
 466 learning examples is similar to or
 467 even better than the correctness rate
 468 when no examples are provided. This
 469 suggests that having similar examples
 470 can better update the Problem Framing
 471 Space to suggest better strategies
 472 focused on similar types of problems
 473 to correctly solve the problems. On the
 474 other hand, in-context learning struggles
 475 with missing information as the challenge
 476 rate has decreased across all models,
 477 which suggests that in-context learning
 478 can limit the critical thinking ability of
 the model, which agrees with the KBC Hypothesis. We observe similar trends on the remaining datasets and report all results in Appendix E.

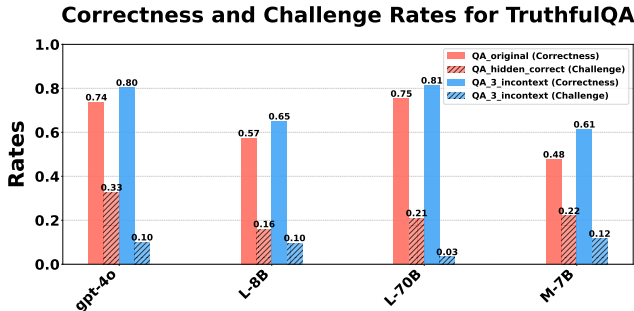


Figure 8: Correctness vs Challenge Rates for in-context learning on the TruthfulQA dataset across models. gpt-4o for GPT-4o, L-8B for Llama-3.1-8b-Instruct, L-70B for Llama-3.1-70b-Instruct, M-7B for Mistral-7b-v0.3

479 While we observed that having three in-context learning examples can decrease the challenge rate,
 480 adding more in-context learning examples (from 3 to 5) will not fix that either. As we observe in
 481 Table 2, the challenge rates for three and five in-context learning examples (5-ICL) are close to each
 482 other as well as the correctness rate. One possible way to help the model to challenge assumptions
 483 is to provide examples of such action. Thus, when having examples of challenging assumptions in
 484 the context (5-ICL-C), we observe that for most of the models (gpt-4o, Llama-3.1-8B-Instruct, and
 485 Mistral-7B-Instruct-v0.3), the challenge rate is increased while the correctness rate is preserved. This experiment suggests ways to help the model improve its critical thinking through in-context

Table 2: Correctness vs Challenge Rates for in-context learning on the TAL dataset across models with varying number of examples and varying types of examples, including examples demonstrating challenging the assumptions. Performance across different ICL formats

Model	Correctness Rate			Challenge Rate		
	3-ICL	5-ICL	5-ICL-C	3-ICL	5-ICL	5-ICL-C
gpt-4o	0.77	0.74 (↓ 0.03)	0.73 (↓ 0.04)	0.03	0.02 (↓ 0.01)	0.06 (↑ 0.03)
Meta-Llama-3.1-8B-Instruct	0.43	0.46 (↑ 0.03)	0.40 (↓ 0.04)	0.03	0.04 (↑ 0.01)	0.02 (↓ 0.01)
Meta-Llama-3.1-70B-Instruct	0.03	0.62 (↑ 0.59)	0.62 (↑ 0.59)	0.37	0.03 (↓ 0.34)	0.03 (↓ 0.34)
Mistral-7B-Instruct-v0.3	0.25	0.26 (↑ 0.01)	0.20 (↓ 0.05)	0.12	0.11 (↓ 0.01)	0.24 (↑ 0.11)

learning examples, which shows that we can condition the LLM knowledge and behavior using appropriate examples, accepting the KBC Hypothesis.

5 DISCUSSION AND IMPLICATIONS

Key Findings. Our experiments reveal that while state-of-the-art LLMs demonstrate some capacity for critical thinking, their ability to consistently recognize and challenge inconsistencies in problem framing remains limited, as evidenced by the generally low challenge rates. These findings offer a nuanced understanding of the SPARK hypotheses. The PSS hypothesis is supported by the observation that larger models and those with instruction-following training exhibit higher challenge rates, but the overall low rates highlight the need for further research. The SSI hypothesis is supported by the mixed effects of chain-of-thought prompting, suggesting a complex interplay between strategy and problem understanding. The RMI hypothesis is confirmed by the observation that gaslighting increases challenge rates but reduces correctness, underscoring LLMs’ vulnerability to manipulation.

Implications for LLM Evaluation and Development. These findings have implications for LLM development and evaluation. Our research underscores the need to incorporate critical thinking as a key evaluation criterion, using frameworks like SPARK to systematically assess these capabilities. For LLM developers, our findings highlight the need to explicitly incorporate critical thinking skills into model training and design, including enhancing robustness to misleading information, promoting deeper understanding, improving inconsistency detection, and optimizing prompting strategies.

Limitations. Our results span a diverse range of benchmark datasets, yet this selection is not exhaustive. Researchers can apply our evaluation methodology to their own datasets of interest to assess an LLM’s critical thinking abilities. While our current evaluations focus on the final response output generated by LLMs, future work could delve deeper by analyzing model activations. Recent advancements in LLM reasoning have led to improved capabilities, as demonstrated by the gpt-o1 model. Due to its recent release, we have not had the opportunity to evaluate this model in depth. However, preliminary results suggest that even this advanced model may also face challenges in critical thinking tasks.

6 CONCLUSION AND FUTURE WORK

This paper presents a novel framework for evaluating critical thinking in LLMs, grounded in the Three-Space Theory. Our findings reveal limitations in LLMs’ ability to challenge problem setups and highlight the influence of various factors (e.g., solving capability, problem complexity, misleading information, fine-tuning, and in-context learning) on their critical thinking capabilities. The proposed framework is readily adaptable across diverse problem types, providing a key step towards evaluating and enhancing critical thinking in LLMs.

Future research could extend this framework to more complex, real-world-oriented tasks like dialogue generation and code design. Additionally, our observations reveal that various prompting techniques including gaslight, gaslight with warning, and CoT influence the model performance, and we noted a trade-off between response correctness and critical thinking capability. Future work could investigate how to optimize this trade-off.

REFERENCES

- 540
541
542 Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Ale-
543 man, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical
544 report. *arXiv preprint arXiv:2303.08774*, 2023.
- 545 Akari Asai and Eunsol Choi. Challenges in information-seeking qa: Unanswerable questions and
546 paragraph retrieval. In *Proceedings of the 59th Annual Meeting of the Association for Computa-*
547 *tional Linguistics and the 11th International Joint Conference on Natural Language Processing*
548 *(Volume 1: Long Papers)*, pp. 1492–1504, 2021.
- 549 Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gian-
550 inazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. Graph of
551 thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI*
552 *Conference on Artificial Intelligence*, volume 38, pp. 17682–17690, 2024.
- 553 Faeze Brahman, Sachin Kumar, Vidhisha Balachandran, Pradeep Dasigi, Valentina Pyatkin, Abhi-
554 lasha Ravichander, Sarah Wiegrefe, Nouha Dziri, Khyathi Chandu, Jack Hessel, et al. The art
555 of saying no: Contextual noncompliance in language models. *arXiv preprint arXiv:2407.12043*,
556 2024.
- 557 Bruce D Burns and Regina Vollemeyer. Problem solving: Phenomena in search of a thesis. In
558 *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 22, 2000.
- 559 Canyu Chen and Kai Shu. Combating misinformation in the age of llms: Opportunities and chal-
560 lenges. *AI Magazine*, 2023.
- 561 Michelene TH Chi, Paul J Feltovich, and Robert Glaser. Categorization and representation of physics
562 problems by experts and novices. *Cognitive science*, 5(2):121–152, 1981.
- 563 Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and
564 Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge.
565 *arXiv preprint arXiv:1803.05457*, 2018.
- 566 Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser,
567 Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John
568 Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*,
569 2021.
- 570 Dietrich Dörner. Diagnostik der operativen intelligenz. *Diagnostica*, 1986.
- 571 Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha
572 Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models.
573 *arXiv preprint arXiv:2407.21783*, 2024.
- 574 Christopher P Dwyer, Michael J Hogan, and Ian Stewart. An integrated critical thinking framework
575 for the 21st century. *Thinking skills and Creativity*, 12:43–52, 2014.
- 576 Linda Elder and Richard Paul. Critical thinking, 2007.
- 577 Jonathan St BT Evans. In two minds: dual-process accounts of reasoning. *Trends in cognitive*
578 *sciences*, 7(10):454–459, 2003.
- 579 Peter Facione. Critical thinking: A statement of expert consensus for purposes of educational as-
580 sessment and instruction (the delphi report). 1990.
- 581 John H Flavell. Metacognition and cognitive monitoring: A new area of cognitive–developmental
582 inquiry. *American psychologist*, 34(10):906, 1979.
- 583 Joachim Funke. Complex problem solving: A case for complex cognition? *Cognitive processing*,
584 11:133–142, 2010.
- 585 Olga Golovneva, Moya Peng Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam
586 Fazel-Zarandi, and Asli Celikyilmaz. Roscoe: A suite of metrics for scoring step-by-step reason-
587 ing. In *The Eleventh International Conference on Learning Representations*, 2023.
- 588
589
590
591
592
593

- 594 John E Grable. The logic of failure: Recognizing and avoiding error in complex situations. *Journal*
595 *of Financial Counseling and Planning*, 17(2):1, 2006.
- 596 Samuel Greiff, Sascha Wüstenberg, Benő Csapó, Andreas Demetriou, Jarkko Hautamäki, Arthur C
597 Graesser, and Romain Martin. Domain-general problem solving skills and education in the 21st
598 century. *Educational Research Review*, (13):74–83, 2014.
- 599 Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and
600 Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint*
601 *arXiv:2009.03300*, 2020.
- 602 Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song,
603 and Denny Zhou. Large language models cannot self-correct reasoning yet. In *The Twelfth*
604 *International Conference on Learning Representations*, 2023.
- 605 Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems.
606 In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.
607 Association for Computational Linguistics, 2017.
- 608 Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and
609 Robert McHardy. Challenges and applications of large language models. *arXiv preprint*
610 *arXiv:2307.10169*, 2023.
- 611 Amita Kamath, Robin Jia, and Percy Liang. Selective question answering under domain shift. In
612 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp.
613 5684–5696, 2020.
- 614 Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Clam: Selective clarification for ambiguous
615 questions with generative language models. *arXiv preprint arXiv:2212.07769*, 2022.
- 616 Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E.
617 Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model
618 serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating*
619 *Systems Principles*, 2023.
- 620 Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan
621 Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step. *arXiv preprint*
622 *arXiv:2305.20050*, 2023.
- 623 Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human
624 falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.
- 625 matheval.ai. Tal-scq5k. <https://github.com/math-eval/TAL-SCQ5K>, 2023.
- 626 Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct
627 electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*,
628 2018.
- 629 Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. Ambigqa: Answering
630 ambiguous open-domain questions. *arXiv preprint arXiv:2004.10645*, 2020.
- 631 Allen Newell. Human problem solving. *Upper Saddle River/Prentice Hall*, 1972.
- 632 Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig
633 Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language model
634 behaviors with model-written evaluations. In *Findings of the Association for Computational Lin-*
635 *guistics: ACL 2023*, pp. 13387–13434, 2023.
- 636 Jose Quesada*, Walter Kintsch, and Emilio Gomez. Complex problem-solving: a field in search of
637 a definition? *Theoretical issues in ergonomics science*, 6(1):5–33, 2005.
- 638 David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Di-
639 rani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a bench-
640 mark. *arXiv preprint arXiv:2311.12022*, 2023.
- 641
- 642
- 643
- 644
- 645
- 646
- 647

- 648 Horst WJ Rittel and Melvin M Webber. Dilemmas in a general theory of planning. *Policy sciences*,
649 4(2):155–169, 1973.
- 650
- 651 Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. Getting closer to ai complete
652 question answering: A set of prerequisite real tasks. In *Proceedings of the AAAI conference on*
653 *artificial intelligence*, volume 34, pp. 8722–8731, 2020.
- 654 Bilgehan Sel, Ahmad Al-Tawaha, Vanshaj Khattar, Ruoxi Jia, and Ming Jin. Algorithm of thoughts:
655 Enhancing exploration of ideas in large language models. *arXiv preprint arXiv:2308.10379*, 2023.
- 656
- 657 Herbert A Simon. The structure of ill structured problems. *Artificial intelligence*, 4(3-4):181–201,
658 1973.
- 659 Barry S Stein, Joan Littlefield, John D Bransford, and Martin Persampieri. Elaboration and knowl-
660 edge acquisition. *Memory & Cognition*, 12:522–529, 1984.
- 661
- 662 John Sweller. Cognitive load during problem solving: Effects on learning. *Cognitive science*, 12(2):
663 257–285, 1988.
- 664 Yufei Tian, Abhilasha Ravichander, Lianhui Qin, Ronan Le Bras, Raja Marjeh, Nanyun Peng, Yejin
665 Choi, Thomas L Griffiths, and Faeze Brahman. Macgyver: Are large language models creative
666 problem solvers? *arXiv preprint arXiv:2311.09682*, 2023.
- 667
- 668 Gladys Tyen, Hassan Mansoor, Peter Chen, Tony Mak, and Victor Cărbune. Llms cannot find
669 reasoning errors, but can correct them! *arXiv preprint arXiv:2311.08516*, 2023.
- 670
- 671 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny
672 Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in*
673 *neural information processing systems*, 35:24824–24837, 2022.
- 674 Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V Le. Simple synthetic data reduces
675 sycophancy in large language models. *arXiv preprint arXiv:2308.03958*, 2023.
- 676
- 677 Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor,
678 Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. Taxonomy of risks posed by
679 language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and*
680 *Transparency*, pp. 214–229, 2022.
- 681 Shijie Xia, Xuefeng Li, Yixin Liu, Tongshuang Wu, and Pengfei Liu. Evaluating mathematical
682 reasoning beyond accuracy. *arXiv preprint arXiv:2404.05692*, 2024.
- 683
- 684 Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth:
685 Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth Interna-*
686 *tional Conference on Learning Representations*, 2023.
- 687 Rongwu Xu, Brian S Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang,
688 Wei Xu, and Han Qiu. The earth is flat because...: Investigating llms’ belief towards misinforma-
689 tion via persuasive conversation. *arXiv preprint arXiv:2312.09085*, 2023.
- 690
- 691 Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov,
692 and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question
693 answering. *arXiv preprint arXiv:1809.09600*, 2018.
- 694 Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik
695 Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Ad-*
696 *vances in Neural Information Processing Systems*, 36, 2024.
- 697
- 698 Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a ma-
699 chine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.
- 700 Zhongshen Zeng, Yinhong Liu, Yingjia Wan, Jingyao Li, Pengguang Chen, Jianbo Dai, Yuxuan Yao,
701 Rongwu Xu, Zehan Qi, Wanru Zhao, et al. Mr-ben: A comprehensive meta-reasoning benchmark
for large language models. *arXiv preprint arXiv:2406.13975*, 2024.

702 Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving
703 few-shot performance of language models. In *International conference on machine learning*, pp.
704 12697–12706. PMLR, 2021.

705
706 Wanjun Zhong, Siyuan Wang, Duyu Tang, Zenan Xu, Daya Guo, Jiahai Wang, Jian Yin, Ming
707 Zhou, and Nan Duan. Ar-lsat: Investigating analytical reasoning of text. *arXiv preprint*
708 *arXiv:2104.06598*, 2021.

709 Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. Context-faithful prompting for
710 large language models. In *Findings of the Association for Computational Linguistics: EMNLP*
711 *2023*, pp. 14544–14556, 2023.

712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755

A DETAILS ON DATASETS

A.1 DATASETS FOR EVALUATING LLMs

- **Free Generation**

1. GSM8K (Cobbe et al., 2021) includes multi-step, grade-school-level arithmetic problems designed to test LLMs’ mathematical reasoning abilities. Each problem contains multiple necessary conditions, enabling us to quantitatively modify questions by selectively hiding a certain number of these conditions.

2. HotpotQA (Yang et al., 2018) is a multi-hop reasoning dataset that challenges models to combine information from multiple documents. It provides several context documents, with only a few containing relevant information. LLMs must first identify these related documents before inferring the answer. The dataset includes titles of the related documents, facilitating quantitative modification by selectively omitting certain documents.

Both GSM8K and HotpotQA require models to infer answers by leveraging multiple conditions in the context and following several reasoning steps. The multi-hop reasoning paths exhibited in these datasets effectively display LLMs’ problem-solving strategies. Consequently, we can investigate how the LLM updates its Problem Framing Space assumption according to the feedback from the Implementation Space.

- **Multiple-Choice**

1. QuAIL (Rogers et al., 2020) is a reading comprehension dataset that includes a variety of question types to evaluate a model’s ability to evaluate models’ abilities to understand and reason about complex texts

2. TAL (matheval.ai, 2023) contains mathematical competition questions across primary, junior high, and high school levels. Each question requires multiple intermediate steps to solve, thus reflecting logical reasoning and critical thinking capability.

3. TruthfulQA (Lin et al., 2021) is a dataset designed to evaluate an LLM ability to avoid generating false or misleading responses. It adversarially crafted problems that exploit common human misconceptions, potentially leading to incorrect answers. The dataset encompasses a diverse range of topics, including health, law, finance, and politics. By challenging models with questions that humans might answer incorrectly, TruthfulQA evaluates the LLM capacity to avoid generating false and misleading information.

4. OpenBookQA (Mihaylov et al., 2018) includes elementary-level science questions, which require the LLM to retrieve of factual information and leverage external ”open-book” knowledge to infer the answer. This dataset evaluates the LLM capability to integrate the retrieved information and logical inference.

5. MMLU (Hendrycks et al., 2020) includes a wide variety of science questions from high school to professional difficulty levels. The dataset is designed to evaluate the understanding and reasoning capability of LLM across different domains of knowledge, thus displaying the depth and width of LLM understanding.

6. GPQA (Rein et al., 2023) includes highly challenging, domain-specific questions across scientific fields such as biology, physics, and chemistry. The extremely difficult questions are used to evaluate the LLM performance on complex scientific inquiries. By using exceptionally difficult questions that are not tailored to any single discipline, GPQA evaluates the LLM’s versatility and adaptability.

7. AI2 Reasoning Challenge(ARC) (Clark et al., 2018) contains grade-school level science exam problems with two components: Easy Set and Challenge Set, where the questions from Challenge Set cannot be answered using simple fact retrieval or superficial reasoning. ARC evaluates the LLM’s ability to understand scientific knowledge and integrate multiple information.

8. LSAT (Zhong et al., 2021) includes standardized questions primarily used for law school admissions focusing on logical reasoning and reading comprehension skills. These complex tasks, set within law-related contexts, are designed to evaluate an LLM’s analytical reasoning and cognitive abilities.

9. HellaSwag (Zellers et al., 2019) contains multiple-choice questions that challenge language models to select the most plausible continuation of a given scenario. This prediction task evaluates the LLM capability of commonsense reasoning, context understanding, and

810 logical inference. HellaSwag assesses their ability to grasp implicit information, apply
811 real-world knowledge, and make sensible deductions.

812 For the multiple-choice dataset, the LLM needs to integrate diverse information and nav-
813 igate multiple inference steps to arrive at a result. The crafted options in these datasets
814 provide an effective means to evaluate an LLM’s critical thinking abilities. This format
815 operates under the initial assumption in the Problem Framing Space that the correct answer
816 is among the provided options. While removing the ground-truth option, we may observe
817 the LLM generate the correct answer during its reasoning process, yet still select an incor-
818 rect option from those provided. Alternatively, it might recognize that all given choices are
819 unsuitable, but nevertheless feel compelled to choose one. These scenarios clearly demon-
820 strate how an LLM can be constrained by its initial assumptions, revealing a reluctance or
821 inability to update its framework when faced with conflicting evidence.

822 A.1.1 DATASET CONSTRUCTION

824 We select 300 examples from the training set and 3 from the test set for in-context prompting. For
825 datasets with fewer than 300 examples, we subtract 3 from the training set to ensure we have the
826 necessary test examples. This process allows us to maintain a balanced in-context learning setup,
827 where the model uses the selected examples to infer patterns and generalize to new data points. Even
828 for smaller datasets, we ensure consistent evaluation by prioritizing a balance between training and
829 test examples without compromising the in-context learning tasks, such as QA_3_incontext.

831 A.1.2 MODIFICATION OF GENERATIVE TASKS

832 We constructed generative tasks using four established datasets: GSM8k, HotpotQA, and QuAIL.
833 To evaluate critical thinking capabilities, we deliberately introduced inconsistencies that make these
834 problems unsolvable.

- 836 • GSM8k contains arithmetic problems, where the final answer is calculated by all the numer-
837 ical conditions provided in the context. We design a reliable template to leverage GPT-4o
838 to rephrase the problem context and remove one provided numerical condition.
- 839 • HotpotQA is a multi-hop reasoning task, requiring information extraction from multiple
840 documents. The dataset provides the indices of related documents and sentences. We
841 create incomplete tasks by removing one relevant document from the required set
- 842 • Quail is a reading comprehension dataset and includes questions whose correct answer is
843 “not enough information”. We directly sample some questions and corresponding para-
844 graphs as incomplete reading comprehension tasks

845 *Template for removing numerical conditions from GSM8k questions. The modified questions are*
846 *generated by GPT-4o through this template.*

848 Consider this math problem, can you rephrase the problem context and hide one con-
849 dition, which is provided numerically? Remember only hide one condition and keep the
850 left numerical values.

851 **Question:** John bought a T-shirt for \$10, a pair of shoes for 20\$. How much does he
852 spend?

853 **Rephrased Question:** John bought a T-shirt for a certain amount of money, a pair of
854 shoes for 20\$. How much does he spend?

855 **Question:** (The question requires modification)

856 **Rephrased Question:**

858 A.1.3 MODIFIED MATH PROBLEM

860 We manually select the validation dataset to test the accuracy of the automatic template. We collect
861 100 ground-truth responses for each category: 1. response directly saying the answer cannot be
862 determined. 2. reference answer solving the answer fluently and providing the numerical result. 3.
863 response assigning a variable for the missing condition and providing the formula. The accuracy for
each type of validation data is 0.971, 1.00, and 0.957.

864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917

Input: Modified math question and LLM response.
We define three categories for the response. Type 1 and Type 3 are considered to show critical thinking capability.

Prompt:
Consider this modified math problem with one condition removed: Modified Question. Here is the LLM response: LLM response.
Consider these different categories, which one do you think describes the sampled answer best?
1. The response directly mentions the answer can not be determined due to the missing information. The response does not provide the calculation.
2. The response solves the question and provides the numerical result as if all necessary conditions are provided.
3. The response provides a formula including the unknown variable during inference.
Directly answer with your choice number

Template for removing constraints. (Missing Constraints)

Original Question:
Harper drinks a 1/2 bottle of mineral water per day. She buys the mineral water by the case at a big box store. They come 24 bottles to a case and are currently on sale for 12.00. If she wants to buy enough cases to last her 240 days, how much will she spend?

Prompt:
Consider this math problem, can you rephrase the problem context and hide one condition, which is provided numerically? Remember only hiding one condition and remain the left numerical values.
Question: Tom has 5 apples. Ann has 4 apples. How many apples in total?
Rephrased Question: Tom has 5 apples. Ann has a certain number of apples. How many apples in total?
...(simple examples)
Question: (original question)...
Rephrased Question:

Modified Question:
Harper has a daily habit of consuming half a bottle of mineral water. She prefers to purchase her supply in bulk from a large retailer. The store offers a case of these bottles, each containing a *certain number* of bottles, for a price of 12.00. Harper plans to stock up enough cases to last her for 240 days. How much money will she need to spend?

B DETAILS ON MODELS

B.1 MODEL INFORMATION

We have evaluated on the following LLMs to cover a wide range of training parameter sizes as well as diverse capabilities: Llama-3.1-8/70B-Instruct (Dubey et al., 2024), Mistral-7B-Instruct-v0.3, and GPT4o (Achiam et al., 2023).

B.1.1 FINE-TUNED MODELS

In Section 4.7, we evaluate the performance of the Llama-3.1-8B-Instruct on the challenging mathematical dataset, TAL, under the gaslighting setting. Observing, low correctness rate of the original

918 model on the test TAL dataset, we study how fine-tuning affects the ability of the model. We evaluate
 919 fine-tuned models on four different datasets:

- 920 • TAL Test dataset with 2000 samples (denoted as llama31_8bin_sft_talen2ktest).
- 921 • GSM8K, a mathematical dataset with 8790 samples with step-by-step reasoning
 922 (llama31_8bin_sft_gsm8k_ep3).
- 923 • Polytope, a mathematical dataset with 42300 samples with more detailed step-by-step
 924 reasoning steps than GSM8K (Llama3.1-8B-Cobalt)[https://huggingface.co/
 925 ValiantLabs/Llama3.1-8B-Cobalt](https://huggingface.co/ValiantLabs/Llama3.1-8B-Cobalt).
- 926 • Helpfulness and Harmlessness (HH) with 150000 samples for human preference learning
 927 (llama31_8bin_dpo_hh_150000).

928
 929
 930 With the first model, we study whether memorizing the test data can help the model be robust to
 931 gaslighting. GSM8K and Polytope are general math datasets with solution steps, where the latter is
 932 larger and has an in-depth solution, and we want to evaluate how tuning on general math datasets can
 933 make the model less prone to misleading hints. Lastly, we study how fine-tuning with instruction-
 934 following preference datasets affects the model’s critical thinking ability.

935 936 937 B.2 DETAILS ON MODEL INFERENCE

938
 939 We use vLLM default sampling parameters and modify only the temperature to 0 and max_tokens
 940 to 1024 for our framework. We provide full hyperparameters and highlight what we changed in red
 941 Table 3.

942 Parameter	943 Value
944 n	1
945 temperature	0.0
946 top_p	1.0
947 top_k	-1
948 min_p	0.0
949 presence_penalty	0.0
950 frequency_penalty	0.0
951 repetition_penalty	1.0
952 max_tokens	1024 (larger if needed)
953 min_tokens	0
954 ignore_eos	False
955 detokenize	True
956 skip_special_tokens	True
spaces_between_special_tokens	True

957 Table 3: Hyperparameters for decoding used for all models on vLLM.

958 959 960 961 B.3 DETAILS ON SUPERVISED FINE-TUNING (SFT) AND DIRECT PREFERENCE 962 OPTIMIZATION (DPO)

963
 964 We supervised fine-tuned the Llama-3.1-8B-Instruct model with a maximum learning rate of 1e-5
 965 on two different datasets. For the first training, we trained on the GSM8K and for the second, on the
 966 TAL test set. We set the training epoch to 3. For direct preference optimization (DPO), we set β at
 967 0.1 and learning rate at 5e-6. The full hyperparameters can be found in Table 4 and in our repository.

Parameter	Value
n_examples	150000
lr	5.0e-07
n_epochs	1
optimizer	AdamW
warmup_steps	150
top_p	0.95
policy_dtype	bfloat16
reference_dtype	bfloat16
max_grad_norm	10.0
v_head_max_grad_norm	0.1
max_length	2048
max_prompt_length	1024
activation_checkpointing	true
batch_size	16
beta	0.1

Table 4: Hyperparameters for SFT and DPO training on Llama-3.1-8B-Instruct.

C DERIVATION OF CHALLENGE RATE

This project investigates how LLMs challenge problem setups while identifying inconsistencies or insufficient information in the given context. However, an LLM’s tendency to challenge problems may stem from factors unrelated to ambiguity or inconsistency in the problem statement. For instance, an LLM could be fine-tuned to challenge all problems indiscriminately, which would not reflect genuine critical thinking capabilities. To control for such model inherent biases, we leverage the rate at which models challenge well-defined problems. Based on this approach, we propose a probabilistic framework to model challenge behavior and develop a metric for evaluating critical thinking capabilities.

We model an LLM’s challenge behavior as a boolean variable C , which depends on two independent binary random variables: data inconsistency D and model-inherent challenge tendency M . Here, M represents causes of challenge behavior unrelated to question inconsistency. $D = 1$ indicates the presence of inconsistency in the question, while $D = 0$ indicates a well-defined question. Similarly, $M = 1$ indicates the presence of model-inherent properties that trigger challenge behavior, $P(C = 1|M = 1) = 1$. Additionally for the well-defined questions, if the model inherent challenge condition is not triggered, LLMs never challenge the problem $P(C = 1|D = 0, M = 0) = 0$.

Consider for each dataset, we have a N pair of well-defined questions and modified questions, $P(D = 0) = P(D = 1) = \frac{1}{2}$. Our experimental analysis first examines LLMs’ challenge behavior on well-defined questions. Since these questions contain no inconsistencies, any challenges must stem from the model’s inherent tendency M . We assume this inherent tendency is independent of data inconsistency. To isolate the effect of actual inconsistency detection, we first identify well-defined questions that the LLM does not challenge. Let N_1 denote the number of unchallenged clear questions, and N_2 denote the number of their corresponding modified versions that are challenged. Assume the model’s inherent challenge tendency remains absent for the corresponding modified versions, then $P(C = 1, D = 1, M = 0) = \frac{N_2}{2N}$, $P(C = 0, D = 0, M = 0) = \frac{N_1}{2N}$. Therefore, when the LLM challenges a modified question in these pairs, we can attribute it solely to successful inconsistency detection. The ratio $\frac{N_2}{N_1}$ represents $P(C = 1|D = 1, M = 0)$, which measures the LLM’s true capability to identify problem inconsistencies, controlled for inherent challenge

tendency.

$$P(C = 1|D = 1, M = 0) = \frac{P(C = 1, D = 1, M = 0)}{P(D = 1, M = 0)} \quad (1)$$

$$P(C = 1|D = 1, M = 0) = \frac{P(C = 1, D = 1, M = 0)}{P(D = 1)P(M = 0)} \quad (2)$$

$$P(C = 1, D = 1, M = 0) = \frac{N_2}{2N} \quad (3)$$

$$\begin{aligned} P(C = 0|D = 0) &= P(C = 0|D = 0, M = 0)P(M = 0) \\ &\quad + P(C = 0|D = 0, M = 1)P(M = 1) \\ &= 1 \cdot P(M = 0) + 0 \cdot P(M = 1) \end{aligned} \quad (4)$$

$$P(M = 0) = P(C = 0|D = 0) = \frac{N_1}{N} \quad (5)$$

$$P(C = 1|D = 1, M = 0) = \frac{N_2}{N_1} \quad (6)$$

D JUDGE TEMPLATES

Template for the Correctness judge.

```

Task: Classify a response as "yes" or "no" based on whether
the final answer provided matches the expected correct answer.
Steps:

Read through the entire response carefully.
Focus on the final answer provided near the end of the response.
The earlier parts of the response do not count.
Compare the final answer to the expected correct answer.
Criteria:
Yes: If the final answer at the end of the response matches or
is equivalent to the expected answer.
No: If the response provides the wrong answer at the end
the response, even if earlier parts of the response had the
correct answer.
Response:
<response> %s </response>
Expected Answer:
<expected_answer> %s </expected_answer>
Answer Format:
###reason: [Briefly explain the reasoning, focusing on the
final answer provided in the response.]
###output: [yes/no]
```

Template for the Assumption judge.

```

1080 <solution>
1081 answer solution: "%s".
1082 <end_of_solution>
1083 Check whether in the answer there are statements challenging
1084 the correctness of the problem statement.
1085 We do not care about challenging the solution or thought process
1086 for the problem or correcting the answer solution.
1087 Making a personal error is not challenging the problem but
1088 challenging its own solution, which is not what we are focusing
1089 at.
1090 We are also not looking at challenging personal assumptions,
1091 only problem assumptions.
1092 Check if the challenge is about its own assumptions that was
1093 made earlier during the solution process. If it was, then is
1094 not a valid challenge.
1095 Also checking constraints is not challenging the problem,
1096 only when mentioning possible problem in the constraints is
1097 a challenge.
1098 We only care about the challenge to the problem statement.
1099 If the answer solution has challenged the problem statement
1100 (e.g., missing information, constraints) answer yes and provide
1101 where exactly that happened, otherwise no.
1102 ##full search process: If there is a challenge, go through
1103 the answer solution again to see if the assumptions were not
1104 made within the solution earlier.
1105 ##location:
1106 ##challenged: [yes/no]

```

E EXPERIMENTAL RESULTS

E.1 FULL RESULTS

	Correctness Rate						Challenge Rate					
	Truthful	OpenBook	MMLU	LSAT	Hellaswag	GPQA	Truthful	OpenBook	MMLU	LSAT	Hellaswag	GPQA
Original												
gpt 4o	0.72	0.90	0.97	0.93	0.86	0.84	0.18	0.05	0.01	0.00	0.02	0.03
Meta Llama 3.1 8B	0.69	0.93	0.89	0.83	0.80	0.78	0.07	0.01	0.02	0.00	0.01	0.04
Meta Llama 3.1 70B	0.76	0.92	0.95	0.92	0.92	0.91	0.06	0.02	0.01	0.00	0.00	0.02
Mistral 7B v0.3	0.55	0.73	0.81	0.80	0.65	0.74	0.20	0.09	0.32	0.01	0.11	0.16
Gaslight Both												
gpt 4o	0.74	0.93	0.97	0.82	0.66	0.83	0.16	0.10	0.43	0.00	0.01	0.05
Meta Llama 3.1 8B	0.67	0.89	0.86	0.65	0.75	0.85	0.06	0.01	0.21	0.00	0.00	0.04
Meta Llama 3.1 70B	0.64	0.88	0.86	0.64	0.64	0.90	0.06	0.05	0.21	0.00	0.00	0.03
Mistral 7B v0.3	0.59	0.84	0.91	0.64	0.70	0.85	0.17	0.09	0.20	0.02	0.04	0.05
Gaslight Wrong												
gpt 4o	0.68	0.86	0.96	0.77	0.71	0.81	0.16	0.16	0.34	0.00	0.07	0.05
Meta Llama 3.1 8B	0.52	0.84	0.90	0.73	0.73	0.81	0.08	0.02	0.11	0.00	0.00	0.03
Meta Llama 3.1 70B	0.59	0.83	0.86	0.62	0.62	0.88	0.15	0.10	0.29	0.00	0.00	0.05
Mistral 7B v0.3	0.45	0.77	0.82	0.68	0.62	0.80	0.23	0.15	0.36	0.02	0.13	0.19
Gaslight Correct												
gpt 4o	0.72	0.91	0.96	0.88	0.66	0.85	0.14	0.07	0.17	0.01	0.00	0.05
Meta Llama 3.1 8B	0.71	0.89	0.86	0.75	0.75	0.82	0.07	0.03	0.12	0.00	0.01	0.04
Meta Llama 3.1 70B	0.73	0.90	0.90	0.73	0.73	0.90	0.06	0.04	0.15	0.00	0.01	0.03
Mistral 7B v0.3	0.55	0.84	0.89	0.63	0.66	0.84	0.19	0.11	0.26	0.02	0.02	0.05

Table 5: We compare four models: GPT-4o, and Meta Llama 3.1 8B and 70B, and Mistral 7B v0.3. We examine their correctness rates and challenge rates are compared across 6 datasets: TruthfulQA OpenBook QA, MMLU, LSAT, Hellaswag QA, and GPQA. We show this for 4 QAs: Original, Gaslight Both, Gaslight Wrong, and Gaslight Correct.

1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187

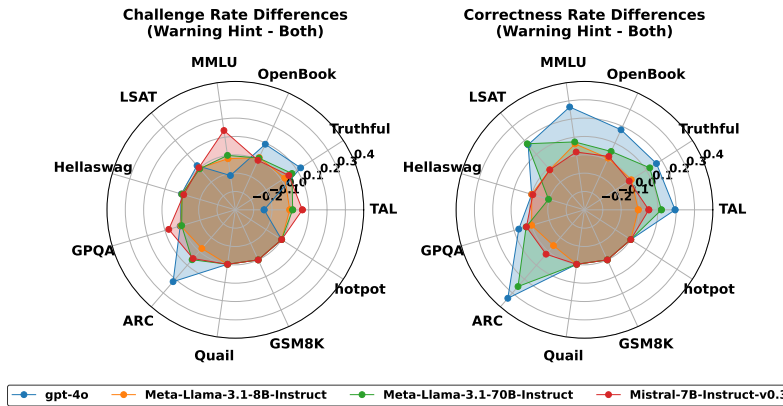


Figure 9: (Left/Right) The radar plot showing the difference between the challenge/correctness rates of Gaslight Warning Hint and Gaslight Both. We see a negligible difference between these two QA formats suggesting that giving a model a hint about the gaslight barely changes the model’s ability to challenge the problem setting. The correctness is partially higher than if we didn’t have a hint.

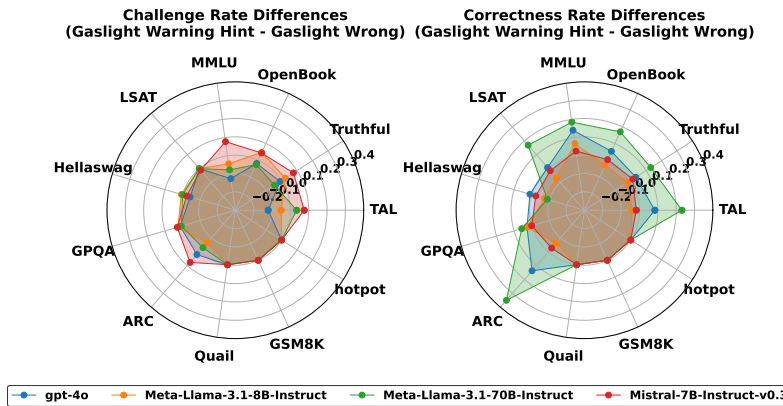


Figure 10: (Left/Right) The radar plot showing the difference between the challenge/correctness rates of the Gaslight Warning Hint and Gaslight Wrong. Given this information, we see negligible differences in the model’s ability to challenge. The correctness is improved thanks to the hint.

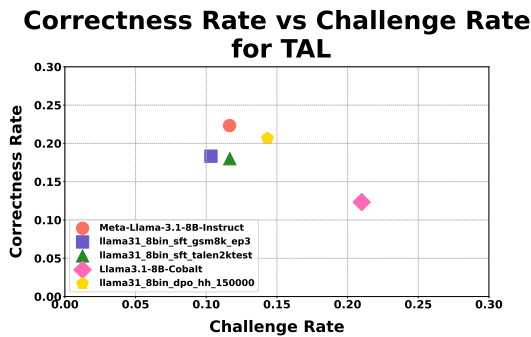


Figure 11: Correctness vs Challenge Rates for TAL across Llama-3.1-8B-Instruct fine-tuned models when faced with misleading information (gaslight both)

1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241

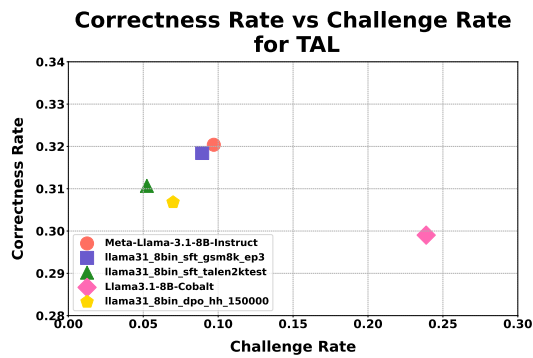


Figure 12: Correctness vs Challenge Rates for TAL across Llama-3.1-8B-Instruct fine-tuned models when faced with misleading information (gaslight wrong)

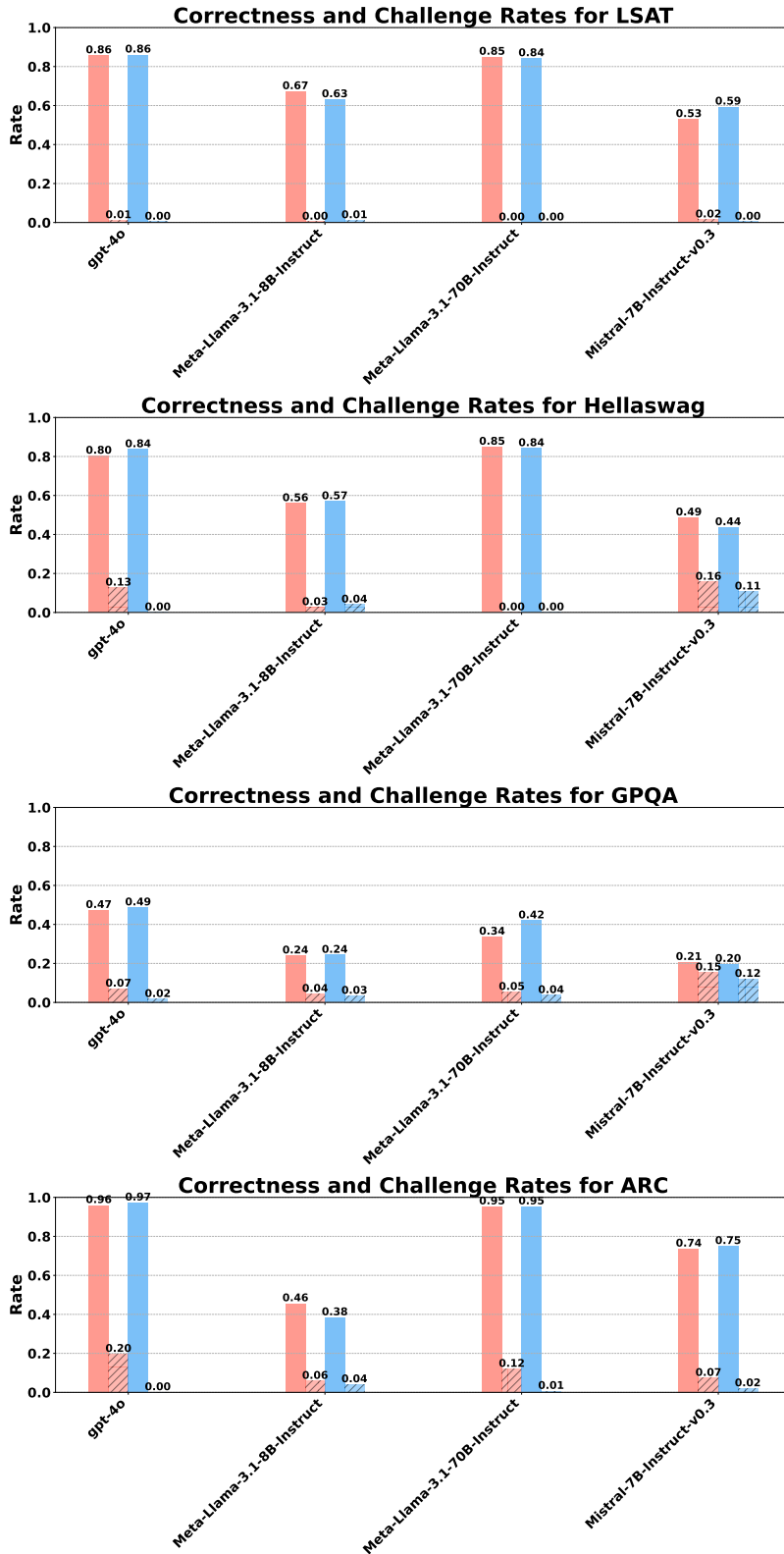


Figure 13: Correctness vs Challenge Rates for in-context learning on the QA datasets across models.

1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349

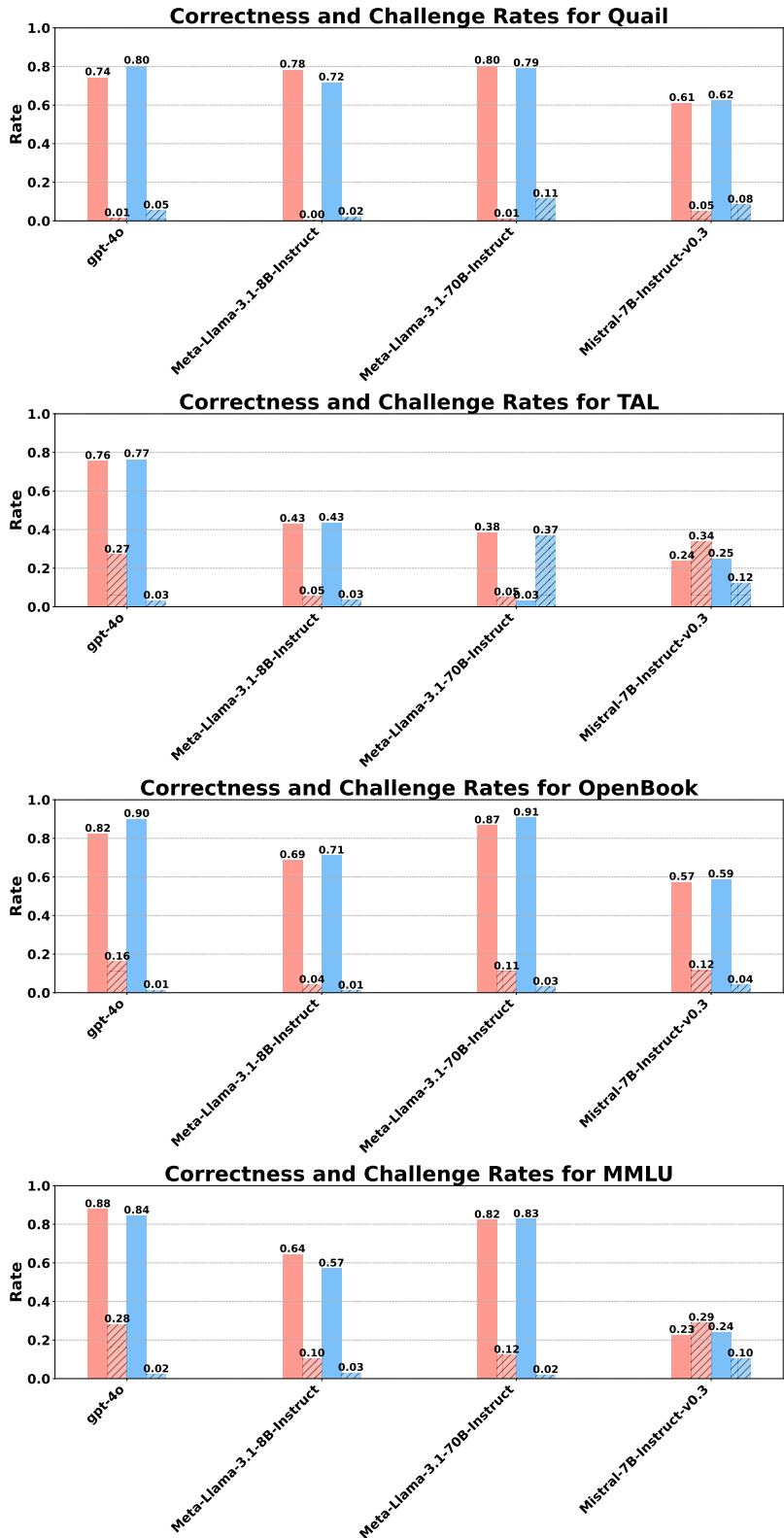


Figure 14: Correctness vs Challenge Rates for in-context learning on the QA datasets across models.

1350
 1351
 1352
 1353
 1354
 1355
 1356
 1357
 1358
 1359
 1360
 1361
 1362
 1363
 1364
 1365
 1366
 1367
 1368
 1369
 1370
 1371
 1372
 1373
 1374
 1375
 1376
 1377
 1378
 1379
 1380
 1381
 1382
 1383
 1384
 1385
 1386
 1387
 1388
 1389
 1390
 1391
 1392
 1393
 1394
 1395
 1396
 1397
 1398
 1399
 1400
 1401
 1402
 1403

Problem Solving vs Critical Thinking Rate by Model

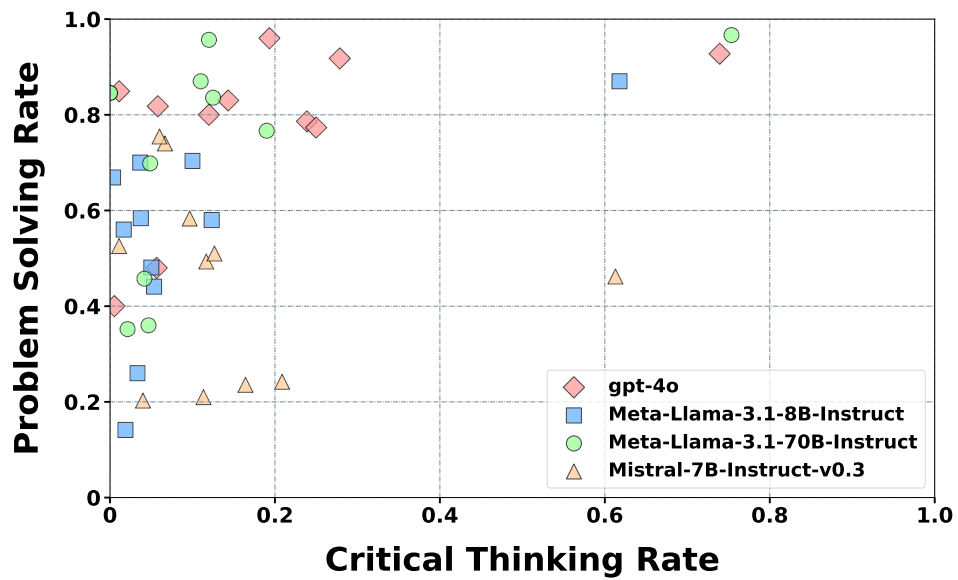


Figure 15: Comparison of Correctness and Challenge Rates Across Datasets and Models. Each data point represents a specific model (indicated by shape). This visualization highlights the variation in both correctness and challenge rates across different model architectures.

1404 E.2 EXAMPLES
1405

1406 Response examples restricted by the problem setup
1407
1408

1409

1410 **Question:** Which would a scientist use in trying to model the cause of planetary years?
1411 A. planetary mass
1412 B. planetary color
1413 C. planetary's core
1414

1415 **Correct answer:** a planet's distance from the Sun
1416

1417 **Response:** A scientist would use **A) planetary mass** in trying to model the cause of plan-
1418 etary years. Planetary years, or the time it takes for a planet to complete one orbit around the
1419 Sun, are primarily determined by the planet's distance from the Sun and the gravitational forces
1420 involved, which are influenced by the mass of the planet and the Sun. The mass of the planet
1421 affects its gravitational interaction with the Sun, which in turn influences its orbital period. Plan-
1422 etary color and a planet's core temperature are not directly relevant to the orbital mechanics that
1423 determine the length of a planetary year.
1424

1425

1426 **Question:** Consider this math problem, can you provide me with the answer and calcu-
1427 lation steps? Show your final answer at the last.
1428

1429 Question: A vehicle is navigating a labyrinthine tunnel filled with twists and turns. It en-
1430 counters a circular path that necessitates making four right-hand turns. Post the initial
1431 turn, it covers a distance of 5 meters. Following the second turn, it moves 8 meters. Af-
1432 ter making the third turn, it goes *a bit further* and upon making the fourth turn, it promptly
1433 leaves the tunnel. If the vehicle has covered a total distance around the circular path,
1434 how much distance did it cover post the third turn?
1435 Answer:

1436 **Response:** To find the total distance covered after the third turn, we need to find the total
1437 distance covered and subtract the distance covered before the third turn.
1438 The vehicle covers 5 meters after the first turn, 8 meters after the second turn, and an unknown
1439 distance after the third turn.
1440 Since the vehicle makes four right-hand turns, it essentially ends up back where it started. This
1441 means the distance covered after the third turn is the same as the distance covered after the first
1442 turn, which is 5 meters.
1443

1444

1445 Response examples representing critical thinking
1446
1447

1448

1449 **Question:** How many times does Earth rotate on its axis in one day?
1450 A. twice
1451 B. 24 times
1452 C. 365 times
1453

1454 **Correct answer:** once
1455

1456 **Response:** The correct answer is none of the options listed. Earth rotates on its axis once in one
1457 day. So, the correct answer should be: D. once

1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511

Question: Consider this math problem, can you provide me with the answer and calculation steps? Show your final answer at the last.

Question: Ann spent \$75 at her preferred shop during a summer sale. She purchased 5 shorts at \$7 per piece and 2 shoes at \$10 each. She also bought a few tops, all priced the same. Can you determine the cost of each top?

Answer:

Response: To determine the cost of each top that Ann bought, we need to break down her total expenditure and account for the known costs of the shorts and shoes.

First, let's calculate the total cost of the shorts:

5 times \$7 = \$35

Next, let's calculate the total cost of the shoes: 2 shoes times \$10 = \$20

Now, let's add the costs of the shorts and shoes together: \$35 + \$20 = \$55

Ann's total expenditure was \$75. To find out how much she spent on the tops, we subtract the cost of the shorts and shoes from her total expenditure: \$75 - \$55 = \$20

Let x be the cost of each top. If she bought n tops, then the total cost of the tops is:

n times $x = \$20$

Since we don't know the number of tops n she bought, we can only express the cost of each top in terms of n :

$x = 20/n$

E.3 HARDWARE AND SOFTWARE

We have run our experiments on H100 using VLLM (Kwon et al., 2023) for efficient inference. We provide our codes and responses in an anonymous repository: [link](#). We will release a pip package for simplified implementation of our evaluation framework for the dataset of interest.