Instruction Boundary: Quantifying Biases in LLM Reasoning under Various Coverage

Anonymous ACL submission

Abstract

Nowadays, automatically generated datasets are increasingly used in LLM reasoning tasks; however, large-scale corpora often contain inherent flaws. For example, a single-choice question may include none or multiple correct options, while true-or-false questions may involve vague or unverifiable statements. We refer to these exceptional answer forms as *sparse labels*. To compare LLMs' ability to recognize various question forms and produce correct answers, we investigate how different instruction formats can either facilitate or mislead LLM reasoning ability. We introduce the concept of *Instruction Boundary*, which systematically analyzes how different levels of prompt coverage—sufficient, redundant, or insufficient—can lead to reasoning biases and performance changes in LLMs. To examine this phenomenon, we design eight experimental settings across five dataset forms. We further propose BIASDETECTOR, a unified framework that quantifies LLMs' ability to identify *sparse labels* under different kinds of *Instruction Boundary* conditions. Evaluations on five mainstream LLMs show that, despite their seemingly high accuracy, substantial reasoning biases persist in many downstream tasks as a direct consequence of prompt coverage. We analyze the impact of these biases and outline possible mitigation strategies. Our findings highlight not only the importance of addressing *sparse labels*, but also the need for developers to recognize and mitigate the risks introduced by *Instruction Boundary*.

1 Introduction

Large Language Models (LLMs), such as GPT-4 (Achiam et al., 2023), have demonstrated advanced reasoning capabilities, ranging from basic tasks like daily conversations, to complex tasks like solving mathematical problems. The performance of LLMs often improves when problems are framed with a set of possible options, as these can guide the model's reasoning process. For instance, multiple-choice questions are typically easier for LLMs to solve than open-ended questions because the provided options offer clues. However, for generated datasets, both question context and options themselves maybe problematic, whereas the task instructions remains the same, leading to biased reasoning and can be further categorized into three main sources: abstention, over-caution and over-confidence, undermining the trustworthiness of LLM reasoning (Lin et al., 2022), (Kadavath and et al., 2022).

As for the problems regarding *sparse labels*, it can be rather under-estimated. The gap between output distribution and accuracy is rarely researched. As shown in the upper part of Figure 1, biased output distribution can lead to seemingly good performance (e.g., answering one specific answer can achieve 50% or higher accuracy), however, as the output distribution doesn't align with label distribution, achieving a high accuracy in name only reflects a limited understanding. To more transparently analyze reasoning biases across related fields, we hereby introduce the concept of *Instruction Boundary*, which refers to the subjective reasoning ability constrained by a prompt's description, task coverage, and contextual guidance. While many works have explored the notion of the Knowledge Boundary, emphasizing whether LLMs should acknowledge their uncertainty and refrain from hallucinations (Lewis et al., 2020; Zhou et al., 2023), the instruction-design bias remains underexplored. Specifically, models can be overly influenced by task framing and exhibit strong obedience to suboptimal instructions (Kojima et al., 2022; Zhao et al., 2021), resulting in brittle or overly confident outputs. This issue is especially critical when prompts provide a biased or incomplete set of options. Consider a case where all options are incorrect—it's far more difficult for the model to reject all of them and respond correctly than to generate an answer from scratch. Similarly, it is essential for LLMs to identify unverifiable options and clearly point them out, or to select multiple useful options

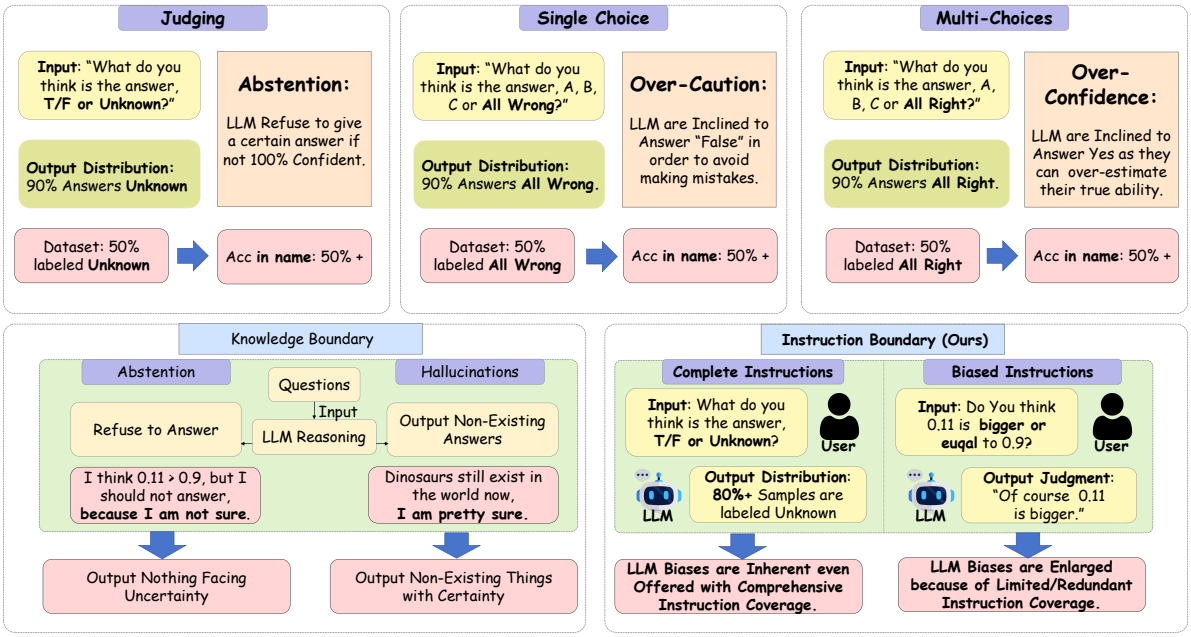

Figure 1: Brief illustrations of potential threat of LLM reasoning biases (Above three). As shown in the picture, even though LLMs can sometimes achieve a 50% or more accuracy in name, however, their understanding towards problems remains limited, which undermines the trustworthiness. And comparisons of previous work Knowledge Boundary and ours definition of *Instruction Boundary* (Below). While previous work mainly focuses on the robustness of LLMs in downstream tasks, we incorporate reasoning biases together with instruction coverage to better reflect real-world scenarios.

when possible.

As a result, we formalize the phenomenon of *Instruction Boundary* regarding five question forms: Judging question that contain (1) True (2) False (3) *Unknown* label, the latter representing unverifiable options; Single-choice questions where questions contain (4) single suitable option (5) multiple suitable options (6) none suitable option. As shown in Table 1. Moreover, we classify labels into two categories according to statistical occurrence, as elaborated in Appendix A. Frequently observed and commonly used labels are classified as *tense labels*, directly determining correctness; whereas no correct option, multiple correct options, and unverifiable options are *sparse labels*. Detailed elaborations are in Section 2.1.

By identifying LLMs' ability to correctly identify *sparse labels* under different instruction settings based on previous work. We argue that understanding the *Instruction Boundary* phenomenon is crucial for developing robust and reliable LLM reasoning. This highlights how the phenomenon in LLMs can cause rigid reasoning errors in generated datasets flaws identifications. Our work provides a holistic review of how task instructions can lead to different biases in LLM reasoning, thereby paving the way not only for future LLM design and development, but also valuable insights for users. In summary, our contributions are three-fold:

▷ We identify and formulate *Instruction Boundary*, a phenomenon in LLM reasoning that leads to biases in LLMs arising from prompt coverage. We propose BIASDETECTOR, a unified framework that quantifies LLMs' ability to determine the reasoning biases under different information provided in the prompt.

▷ By categorizing labels as tense and sparse based on their frequency of occurrence, we examine the flaws generated datasets may have regarding unusual question forms. Through systematic analyses across different experimental settings, we establish evaluation metrics that precisely characterize the biases introduced by question forms and *Instruction Boundary*. These insights help us better understand the paradigm of LLM reasoning.

▷ By comparing BIASDETECTOR with diverse related benchmarks, we assess how principles from related areas generalize, and analyze how much information should be given to LLMs, providing a new perspective for deepening our understanding of LLM trustworthiness.

Table 1: Datasets categorized by prompt form, showing task type, dataset, sample sizes, and data forms.

Type	Dataset	Samples	Form
Judging	FLD (Morishita et al., 2024)	600	True/ False
	FOLIO (Han et al., 2022)	640	**Unknown**
Multiple Choices	ScienceQA (Lu et al., 2022b)	150	Single-select **All Wrong**
	MMLU-Pro (Wang et al., 2024b)	300	
	MMLU (Hendrycks et al., 2021a)	50	
	CommonsenseQA (Talmor et al., 2019)	50	
	TruthfulQA (Lin et al., 2022)	50	
	MultiRC (Khashabi et al., 2018)	150	**Multiple-select**

2 Evaluation Framework

2.1 Problem Formulation

Phenomenon Settings *Instruction Boundary* covers three situations: (1) **Sufficient Instructions** with full task coverage, where reasoning biases arise from the LLM's internal tendencies. (2) **Redundant Instructions** with complete but distracting information, which can potentially misguide LLMs. (3) **Insufficient Instructions** that leave out specific critical options. *Instruction Boundary* can be applied to a wide variety of experiment settings. As shown in Table 2, we extract eight different settings that broadly cover multiple aspects. Refer to Appendix C for detailed information.

Label Categorization We classify labels into *tense* and *sparse*. Tense labels are commonly seen and tend to yield higher robustness and prediction likelihood for LLMs; Sparse labels are infrequent and difficult for LLMs to consider without explicit cues. However, as generated datasets are more and more commonly used nowadays, researching *sparse labels* caused by datasets generation problems can be essential. In our work, single-choice and normal True-or-False are tense labels, whereas unknown label, multiple suitable options, and no suitable option are categorized as sparse labels. As shown in Table 1, *sparse labels* are marked darker. Detailed statistics of datasets and labels can be found in Appendix A.

2.2 Proposed Framework

Datasets & Pre-processing Based on reasoning tasks, we categorize the datasets into twofold based on the different forms: (1) True-False-Unknown Dataset: FLD (Morishita et al., 2024), and also the FOLIO (Han et al., 2022) (2) Multiple-choice questions, containing samples from ScienceQA (Lu et al., 2022b), MultiRC (Khashabi et al., 2018), MMLU (Hendrycks et al., 2021a), MMLU-Pro+ (Wang et al., 2024b), CommonsenseQA (Talmor et al., 2019), TruthfulQA (Lin et al., 2022). Firstly, we pre-process all the three forms of datasets to uniformly fit into the framework: (1) For T-F-Unknown datasets like FLD (Morishita et al., 2024) and FOLIO (Han et al., 2022), labels that are objectively unverifiable are marked as *Unknown* serving as *sparse labels*, which contrasts against others that are explicitly labeled as either *True* or *False*. (2) Similarly, for multiple-choice questions in both datasets, we extract 50 samples per dataset from CommonsenseQA, MMLU and TruthfulQA, representing samples with one suitable option. Meanwhile, we extract 150 samples that contains five options in total with one correct option, by masking the right one can we finally get samples with none suitable option. For multiple options case, we directly extract 150 samples from MultiRC dataset. Each dataset contains an equal distribution regarding question forms, with half containing *sparse labels* and the other *tense labels*. Refer to Appendix A for detailed information about datasets.

Metrics First of all, we introduce the very basic concept: (1) **Accuracy** of tasks, representing the basic reasoning ability in settings regarding label output matching, used in all settings:

$$Acc = \frac{1}{|D|}\sum_{i=1}^{|D|}\mathbb{1}\big(y^i = \hat{y}^i\big)$$

To quantify whether the LLM reasoning is robust or not, we use the following three metrics. We first run one turn under *Vanilla Scenario* setting, then we apply experimental settings prompt, details are explained in Table 2. Based on outcomes, we define three metrics: **Generalization Rate (GR)** and **Stability Rate (SR)**, **Robustness Rate (RR)** as below:

$$GR = \frac{1}{|D|}\sum_{i=1}^{|D|}\mathbb{1}\big(y^i = \hat{y}_b \ \wedge \ y^i \in D_S\big)$$

$$SR = \frac{1}{|D|}\sum_{y^i \in D_T}\mathbb{1}\big(y^i = \hat{y}_b^i\big)$$

$$RR = \big|Acc_v - Acc_b\big|, \quad D_S \cup D_T = D$$

For y^i indicates the true label, $\hat{y}^i$ is the predicted label, for $\hat{y}_b^i$ denotes the biased predicted label from one of experimental settings. L refers to the total label categories set and D_T, D_S means the subset of datasets where all samples have tense or sparse Labels. RR indicates the accuracy change of samples

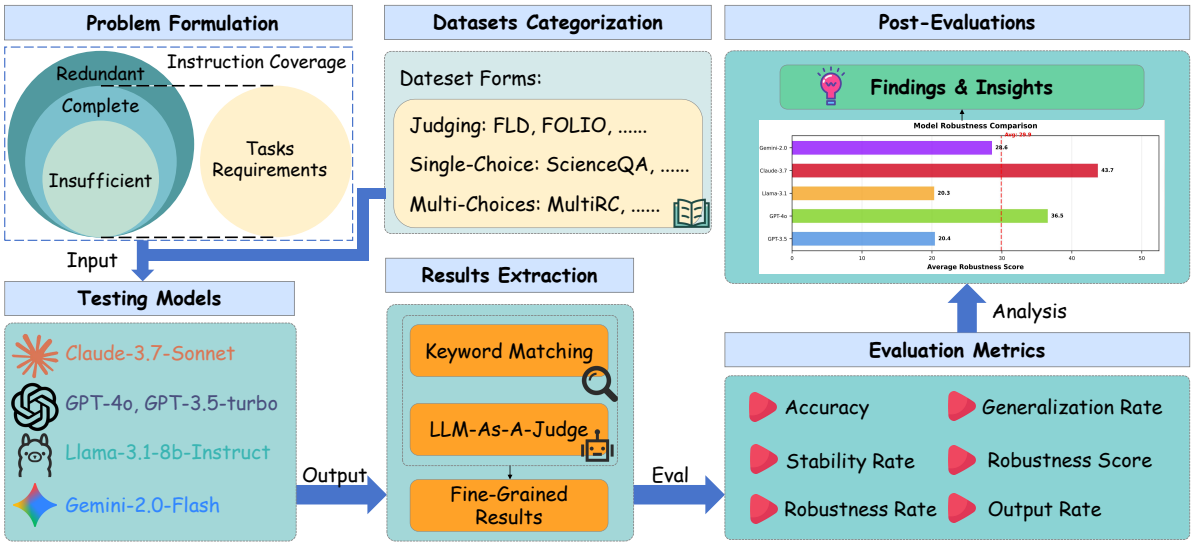

Figure 2: General pipeline of our proposed framework BIASDETECTOR. First of all, we decompose the phenomenon by settings and define problem types regarding different aspect of *Instruction Boundary*. Then we apply five LLMs from different brands to generate biased reasoning under these settings. Finally by testing their robustness and accuracy regarding different aspects, we discover problems and thus offer suggestions for future LLM improvements.

under Complete and Insufficient / Redundant Instruction settings; GR represents the model successfully identify and output the *sparse label*. SR represents the accuracy regarding *tense labels*, which can serve as a comparison with GR to see whether the reasoning ability is aligned or not. These three metrics help better quantify the biases post by *Instruction Boundary* phenomenon. Thus, for common scenarios with sufficient instructions, we can calculate the **Robustness Score (RS)**, denoting the harmonic mean of SR and GR. We measure LLM reasoning biases by calculating **Output Rate (OR)** as well, which is expressed as:

$$OR = \frac{1}{|D|} \sum_{i=1}^{|D|} \mathbb{1}(\hat{y}^i \in D)$$

$$RS = \frac{1}{N} \sum_{n=1}^{N} \frac{2\, SR_n \cdot GR_n}{SR_n + GR_n}, \quad n = 1, 2$$

N denotes the number of dataset forms within our experiments. The accuracy gap between sparse and tense label would further depress *RS*, indicating misaligned reasoning ability; whereas higher RS indicates better alignment between robustness and generalization. By measuring both accuracy and output preferences, we can understand how LLMs handle these problems

3 Experiments

3.1 Experimental Setup

Models As discussed in (Ling et al., 2025), even though higher parameter sizes in LLMs may lead to enhanced reasoning capabilities, models are more willing to have *Abstention* and thus output unknown even if the sample has a True/False label. In this study, we evaluate both closed-source and open-source LLMs across a wide range of brands, including Claude-3-7-Sonnet (Anthropic, 2024), Gemini-2.0-Flash (DeepMind, 2024), LLaMA-3.1-8B-Instruct (Touvron et al., 2023b), GPT-3.5-turbo and GPT-4o (OpenAI, 2023, 2025).

3.2 Findings of Complete Instructions

As mentioned previously, we do not rule out the possibility of LLMs themselves are inherently biased thinkers. Notably, LLMs show consistent bias patterns under both *Vanilla Scenario* and *Multi-turn Dialogue* settings.

Vanilla Scenario As illustrated in the right two panels of Figure 3, LLMs still exhibit explicit and unified bias even when presented with all options: they frequently output *Unknown* and *True*, for multiple-choice questions, outputting one single choice takes up a significant amount, indicating a strong preference for specific labels. Meanwhile, Table 3 show that the accuracy gap between different label types (SR and GR can be extremely large, further proving LLMs' ability misalignment in different question forms.

Table 2: *Instruction Boundary* settings: classification of biased prompts and their corresponding tasks, with descriptions illustrating how different prompt coverage impacts LLMs' reasoning biases. *Instruction Boundary (IB)* is divided into three main categories: (1) Sufficient (2) Redundant, and (3) Insufficient Instructions. Each category is further divided into sub-categories. One setting is carried out separately: **Disturbing Miscellany** is based on MMLU-Pro (Wang et al., 2024b) with 10 candidate options.

Sufficient-Instructions	**Paradigm**	**Examples**
Vanilla Scenario	Instruction coverage is comprehensive, providing detailed and unbiased guidance as the empirical baseline input.	Which is the answer? True: Conclusion is proved. False: Conclusion is disproved. Unknown: Conclusion cannot be determined.
Multi-turn Dialogue	Querying more than once triggers self-reflection, which may increase or reduce reasoning bias.	Which is the answer? T:..., F:..., or Unknown:... Also show your reasoning process. (Encourages further reasoning with follow-up input.)
Redundant-Instructions	**Paradigm**	**Examples**
Conformity	LLMs may exhibit bias toward labels (tense or sparse) that are explicitly mentioned or implied in the prompt.	Which is the answer? T:..., F:..., or Unknown:... The answer is .... (A specific tense or sparse label is suggested.)
Disturbing Miscellany	Additional distracting options are added to the prompt, which increases deviation from the intended answer.	Which is the answer? T:..., F:..., Unknown:..., All above are true:..., All above are false:... (Contains extra unrelated distracting information.)
Few-shot Learning	LLMs may be biased toward outputting labels that are present in the few-shot examples, even if inappropriate.	Which is the answer? T:..., F:..., or Unknown:.... Here is an example to learn from. (Few-shot examples with specific tense or sparse labels may affect output.)
Insufficient-Instructions	**Paradigm**	**Examples**
Missing Choices	Some correct options are omitted from the prompt, causing bias toward the available choices.	Which is the answer? T:..., F:..., Unknown:..., All above are true:... (At least one correct option omitted.)
Vagueness	Prompts provide full coverage but lack detailed guidance, making the reasoning process ambiguous.	Which is the answer? True, False, or Unknown. (Details are omitted.)
Prompt-Polishing	Asking LLMs to polish prompts to improve understanding, which may exacerbate bias if the input is limited.	Which is the answer? T:..., F:..., or Unknown:... (Prompt polished up by LLMs.)

Insight 1: LLMs exhibit substantial biases even all label types are informed, demonstrating that fairness in outputs is inherently limited.

Multi-turn Dialogue We use a multi-turn dialogue protocol whereby the model first generates an answer and then is prompted to review and refine its response in subsequent turns.

Insight 2: As depicted in Table 3, while multi-turn prompting can narrow the gap between Stability Rate (SR) and Generalization Rate (GR), it yields only modest improvements, indicating that the inherent biases remain largely unmitigated.

Sub-conclusions Even with complete option coverage and explicit guidance, LLMs continue to exhibit pronounced and consistent reasoning biases at a global scale, suggesting that **most LLMs remain strongly influenced by their own inclinations**.

3.3 Findings of Redundant Instructions

Learning by imitations is commonly adopted in LLM reasoning applications, inspiring models to think given the clues. However, the quality of examples/guidance within prompt can play an important role in reasoning performance and fairness. However, to the best of our knowledge, we are the first to

Table 3: Main Experimental Results among six LLMs across seven settings except for *Disturbing Miscellany*, which is illustrated later in Fig 3. The superscript number indicates different question forms: 1 means Judging datasets. and 2 means multiple-choice datasets. GR^2 and GR_3 represents the accuracy on samples' labeled "none suitable option" and "multiple suitable options" respectively.

Model	Setting	Acc_b^1	SR^1	GR^1	Acc_b^2	SR^2	GR^2	GR^3
GPT-3.5-turbo	Vanilla	48.75	21.47	76.03	24.67	62.00	12.00	0.00
	Missing Choices	34.76	61.04	8.48	25.56	64.00	12.67	0.00
	Few-shot Learning	50.00	**99.38**	0.63	29.11	65.33	22.00	0.00
	Vagueness	38.32	23.17	50.38	24.70	73.30	0.70	0.00
	Prompt-polishing	34.30	9.16	59.44	21.33	58.67	5.33	0.00
	Multi-turn	36.14	31.51	40.78	33.56	40.67	60.00	0.00
	Conformity	50.00	0.00	**100.00**	37.11	11.33	**100.00**	0.00
GPT-4o	Vanilla	45.41	63.49	27.32	39.33	89.33	28.67	0.00
	Missing Choices	40.39	38.09	47.69	40.89	90.00	32.67	0.00
	Few-shot Learning	56.12	80.73	31.50	44.00	**98.00**	34.00	0.00
	Vagueness	41.01	12.28	69.75	32.70	93.30	4.70	0.00
	Prompt-polishing	48.86	60.50	37.21	37.11	89.33	19.33	2.67
	Multi-turn	40.07	29.53	50.62	18.89	16.67	40.00	0.00
	Conformity	56.92	20.60	93.24	44.00	83.33	48.67	0.00
Llama-3.1-8b-Instruct	Vanilla	51.87	17.92	85.82	20.67	55.33	4.67	2.00
	Missing Choices	47.91	8.67	86.67	22.44	59.33	5.33	**2.67**
	Few-shot Learning	46.28	77.40	15.15	38.00	50.67	63.33	0.67
	Vagueness	29.21	0.32	58.11	25.60	75.30	0.70	0.70
	Prompt-polishing	42.58	50.28	34.88	7.78	13.33	5.33	4.67
	Multi-turn	38.81	57.70	19.93	15.33	36.00	8.67	1.33
	Conformity	50.00	0.00	100.00	38.00	50.67	60.00	3.33
Claude-3.7-Sonnet	Vanilla	46.23	40.13	52.33	7.78	12.67	7.33	3.33
	Missing Choices	53.19	26.32	80.07	9.78	14.67	11.33	3.33
	Few-shot Learning	54.10	54.09	54.11	25.11	52.67	21.33	1.33
	Vagueness	51.20	45.34	57.07	30.20	88.70	2.00	0.00
	Prompt-polishing	**57.21**	84.06	27.66	2.67	1.33	5.33	1.33
	Multi-turn	45.73	55.48	35.98	14.00	16.67	24.67	0.67
	Conformity	55.10	18.70	91.49	3.78	1.33	7.33	2.67
Gemini-2.0-Flash	Vanilla	51.81	36.99	66.62	35.00	84.85	18.18	2.94
	Missing Choices	45.79	31.30	60.28	36.00	87.88	21.21	0.00
	Few-shot Learning	54.51	88.23	18.31	42.67	87.33	40.67	0.00
	Vagueness	40.73	19.49	61.96	31.60	90.70	4.00	0.00
	Prompt-polishing	35.40	0.50	70.31	10.44	18.67	7.33	**5.33**
	Multi-turn	40.84	39.81	41.87	36.00	60.67	45.33	2.00
	Conformity	51.20	4.68	97.70	**50.00**	65.33	84.67	0.00

compare *Few-Shot Learning* with *Conformity*. The former contains a representative example for LLM to follow and study from; whereas the latter directly offers the model a specific label to influence its reasoning. Besides, we also analysis *Disturbing Miscellany* settings, which dive into how redundant options would disturb the performance.

Few-shot Learning In the setting, we only add samples with *sparse labels*. we can tell from the table 3, compared to *Vanilla* settings, few-shot learning always lead to a accuracy gain, however, the value of SR and GR can be largely changed. Details are explained in Appendix F.

Insight 3: Figure 4 shows that example labels in Few-shot learning can trigger huge biased output. LLM performance gain may be largely attributed to a "bias shift" rather than "performance gaining".

Conformity LLM reasoning biases can be extremely higher when sole labels are offered. Figure 3

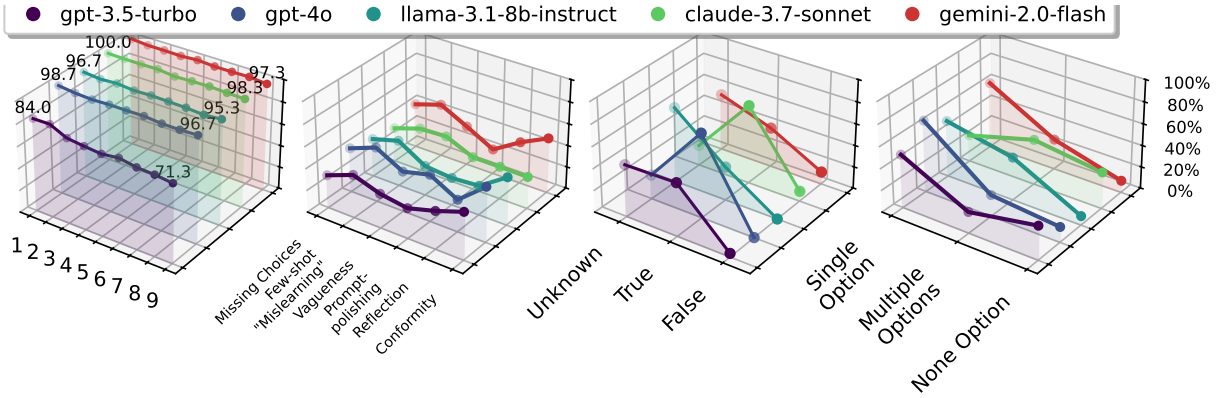

Figure 3: Results of *Accuracy* of six LLMs experiments under *Disturbing Miscellany* setting, the accuracy is dropping as the number of options is increasing according to the x-axis number of options (left) and general *Accuracy* among three datasets forms (middle), and *Label Output Distribution (OR)* under *Vanilla Scenario* settings (right). Standardized with the scale of the rightmost figure as the reference.

indicates that even though accuracy doesn't change a lot, however, the gap between *SR* and *AR* is even higher compared with previous settings. Details of output distribution are shown in Fig 4, charts (2)(3) demonstrate an extremely higher output distribution shift in most mainstream LLMs whether the label is often or rarely-seen ones. However, there are exceptional models like GPT-3.5-turbo that demonstrates *anti-conformity*, outputting *unknown* at dominate ratio no matter which labeled examples are offered.

Insight 4: LLMs demonstrate much higher conformity regarding sparse labels like *Unknown*, compared with tense labels like *True* or *False*. Meanwhile, for unstable models like LLaMA-3.1-8B-Instruct, accuracy can be seriously influenced by one single sentence: model output distribution is mostly determined by conformity and what it is told, inducing risks on LLM Judge usage.

Disturbing Miscellany In this experiment, we research the accuracy change regarding redundancy choices offered to LLMs. We choose MMLU-Pro (Wang et al., 2024b), a complex single-choice question dataset which consists one true option and up to nine false options per sample. In our experiment, we save **N** numbers of options to see how the redundant and false prompt influence LLM reasoning performance, and **N** ranges from one to nine.

Insight 5: According to the left chart of Fig 3, most LLMs show a slight drop in accuracy, however, except for the elder ones like GPT-3.5-turbo demonstrates a higher decline, others still maintain high precision and robustness regarding redundant options added, indicating LLMs nowadays are not easily misguided offered as long as it is not treated as a rule to follow like Conformity.

Sub-conclusions *Disturbing Miscellany* settings prove that the ability to get rid of disturbance has been strong for up-to-dated models, and strong conformity effects does not come from ability shortage. For the three settings above, *Conformity* induces the most serious biases, *Few-shot Learning* ranks the second, and *Disturbing Miscellany* the last.

3.4 Findings of Insufficient Instructions

Prompt Polishing Users like to polished up prompts with LLMs to achieve better results. We observe a widely accuracy drop among all models, ranging from 5% to 30%. Meanwhile, most LLMs demonstrates a even larger gap between Generalization Rate (GR) and Stability Rate (SR), and show distinct difference of output distribution compared with *Vanilla Scenario*.

Insight 6: Prompt polishing is likely to introducing instability in model outputs, thereby increasing reasoning biases and potential risks.

Vagueness Omitting detailed guidance and information of each option, leading to a great drop in Robustness Score (RS) as shown in Table 7 and causes accuracy to decline by tens of percentage points, exceeding model-wise performance differences.

Insight 7: The primary bottleneck in downstream tasks is the comprehension gap of the problem, not its problem-solving capability once understood; excessive prompt freedom often degrades performance.

Missing Choices Leaving out specific choices and requiring models to output it with self-awareness can be hard. Figure 3 shows an overall accuracy decline across settings; LLMs often resort to abstention or produce hallucinations under challenging prompts.

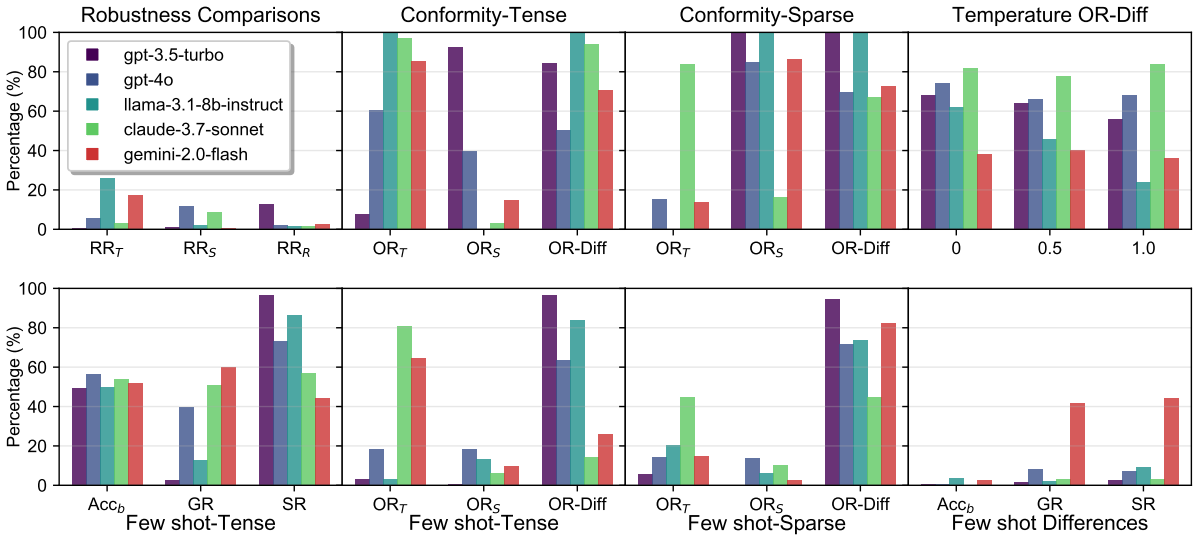

Figure 4: Comparisons of *Conformity*, *Few-shot* and *Disturbing Miscellany* settings. The **Generalization Rate (GR)**, accuracy difference between redundant options and vanilla setting is depicted. For the two in the middle are under conformity setting **Output Rate (OR)** and difference of two kinds among five LLMs experiments under conformity evaluations. Metrics details are elaborated in Section 2.2. The eightieth chart is Acc_b, GR and SR difference of *Few-shot setting* between tense and sparse label setting, from which we can tell Gemini-2.0-flash is largely influenced by Few-shot example types.

Insight 8: When options information is excessively constrained, LLMs occasionally infer beyond provided options but still suffer a 5–20% accuracy loss. Notably, aside from older models like GPT-3.5-turbo, newer LLMs exhibit high abstention rates, suggesting independent inference at the cost of output stability.

Sub-conclusions Even though all insufficient-instruction settings incur accuracy declines, LLMs retain resilience against prompt-induced risks.

4 Discussion

We identified three types of biases sources: complete, redundant and insufficient instructions. Complete instructions can result in biases even though detailed and full guidance are offered; On the other hand, redundant and insufficient instructions make the biases even larger by including confusing elements in the prompt. Our case studies illustrate these biases in Appendix M. Although clear evidence shows that LLM still lacks the ability to give fair and justice output regarding whether the scenario is familiar or not, however, sometimes sticking to often-seen *tense labels* can be of benefit in a number of tasks, as it always means higher attack-resistance. For all three settings, huge biases are detected. Thus we offer three suggestions to help:

▷ **Users' strong intentions may cause conformity. Simple toolkits can be developed to better wrap user considerations in prompts as an option rather than an order to follow.**

▷ **More focus should be placed on the faithfulness of LLM reasoning, as they are inner biases thinkers regardless of objective situations.**

▷ **Be cautious about LLM-based prompt polishing, and ensure it contains the necessary information to help LLMs wrap things up.**

5 Conclusion

In this paper, we formalized *Instruction Boundary*, a phenomenon resulting in huge and under-estimated reasoning biases because of limited/redundant instruction coverage, or even inner reasoning biases resulting from LLM attributes instead of prompt coverage. By testing LLMs under our proposed framework, we show that biases caused by flawed input coverage can substantially affect tasks accuracy and reliability, which brings about tremendous potential influence when it comes to various applications as it is often the case that user prompts can be flawed and biased. Our work prove that even though related research focusing on Knowledge Boundary, where training model to learn when to refuse to answer facing vague understanding, their answering can also be even significantly affected by *Instruction Boundary*, as the performance are determined by commonly used modifications on prompt and its coverage. How to achieve unbiased LLM reasoning still remains a problem to be solved in the future.

5.1 Limitations

Although we address three categories of prompt coverage problems, our work does not consider mainstream workflows such as LLM agents or RAG systems. Moreover, our study focuses primarily on language-related tasks, leaving vision-related challenges outside the scope of this paper. Besides, for real-life applications, user prompt forms may be varied and noisy, leaving the final performance instable. Unified analysis regarding different elaboration forms is valuable but too large to be examined.

5.2 Ethical Considerations

All five datasets are **publicly released for non-commercial research**. FLD is entirely synthetic; FOLIO, ScienceQA, MMLU-Pro, and MMLU derive from open educational resources and contain no personally identifiable information. We abide by each dataset's license (CC BY-SA 4.0 for FOLIO, CC BY-NC-SA 4.0 for ScienceQA, Apache 2.0 for MMLU-Pro, CDLA-Permissive 2.0 for MMLU) and distribute only aggregate statistics. No additional human data were collected, so the study does not require IRB review; environmental impact is minimal because we perform inference only, with no model training.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anonymous. 2023. Investigating data contamination in modern benchmarks for large language models. *arXiv preprint*.

Anthropic. 2024. Introducing claude 3.7 sonnet. https://www.anthropic.com/news/claude-3-7. Accessed: 2025-09-15.

Samee Arif, Zohaib Khan, Agha Ali Raza, and Awais Athar. 2024. With a grain of salt: Are llms fair across social dimensions? *arXiv preprint arXiv:2410.12499*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, and 32 others. 2022. Constitutional ai: Harmlessness from ai feedback. *Preprint*, arXiv:2212.08073.

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. 2025. Hallulens: Llm hallucination benchmark. *Preprint*, arXiv:2504.17550.

Bowen Cao, Deng Cai, Zhisong Zhang, Yuexian Zou, and Wai Lam. 2024. On the worst prompt performance of large language models. *arXiv preprint arXiv:2406.10248*.

Muhammad Umer Cheema, Jiyang Jin, Xiang Xie, Yitian Shao, Yifan Hou, Yuzhong Huang, Minghui Zhang, Qi Zhang, and Jiliang Sun. 2024. Trustllm: A benchmark for trustworthiness of large language models. *arXiv preprint arXiv:2403.09028*.

Khaoula Chehbouni, Mohammed Haddou, Jackie Chi Kit Cheung, and Golnoosh Farnadi. 2025. Neither valid nor reliable? investigating the use of llms as judges. *Preprint*, arXiv:2508.18076.

Karl Cobbe, Vinay Kosaraju, Mohammad Bavarian, Mark Chen, and et al. 2021. Training verifiers to solve grade school math word problems. *arXiv preprint arXiv:2104.03253*.

Google DeepMind. 2024. Gemini 2.0 flash. https://deepmind.google/technologies/gemini/flash. Accessed: 2025-09-15.

Deep Ganguli, Amanda Askell, Yuntao Bai, and et al. 2022. Red teaming language models with language models. *arXiv preprint arXiv:2210.07237*.

Shunyu Gao, Swaroop Mishra, Daniel Huang, Thomas Liu, Saket Mittal, and et al. 2023. Pal: Program-aided language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Gracjan Góral and Emilia Wiśnios. 2024. When all options are wrong: Evaluating large language model robustness with incorrect multiple-choice options. *arXiv preprint arXiv:2409.00113*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. *arXiv preprint arXiv:2310.14566*.

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*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, and et al. 2021b. Measuring mathematical problem solving with the math dataset. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Tiancheng Hu, Yara Kyrychenko, and Jon Roozenbeek. 2025. Generative language models exhibit social identity biases. *Nature Computational Science*. Published online December 2024.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. 2023. Llama guard: Llm-based input-output safeguard for human-ai conversations. *Preprint*, arXiv:2312.06674.

Saurav Kadavath and et al. 2022. Language models struggle to generalize across prompting formats. *arXiv preprint arXiv:2205.00415*.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Chaoyu Li, Eun Woo Im, and Pooyan Fazli. 2025. Vidhalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. *arXiv preprint arXiv:2412.03735*.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. Halueval: A large-scale hallucination evaluation benchmark for large language models. *Preprint*, arXiv:2305.11747.

Moxin Li, Yong Zhao, Yang Deng, Wenxuan Zhang, Shuaiyi Li, Wenya Xie, See-Kiong Ng, and Tat-Seng Chua. 2024a. Knowledge boundary of large language models: A survey. *arXiv preprint arXiv:2412.12472*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *Proceedings of EMNLP 2023*. ArXiv:2305.10355.

Yifei Li, Xiang Yue, Zeyi Liao, and Huan Sun. 2024b. Attributionbench: How hard is automatic attribution evaluation? In *Findings of the ACL 2024*, pages 14919–14935.

Percy Liang, Rishi Bommasani, Kenny Lee, and et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of ACL 2022*. ArXiv:2109.07958.

Chen Ling, Xujiang Zhao, Xuchao Zhang, Wei Cheng, Yanchi Liu, Yiyou Sun, Mika Oishi, Takao Osaki, Katsushi Matsuda, Jie Ji, Guangji Bai, Liang Zhao, and Haifeng Chen. 2024. Uncertainty quantification for in-context learning of large language models. *Preprint*, arXiv:2402.10189.

Zipeng Ling, Yuehao Tang, Shuliang Liu, Junqi Yang, Shenghong Fu, Chen Huang, Kejia Huang, Yao Wan, Zhichao Hou, and Xuming Hu. 2025. Wakenllm: Evaluating reasoning potential and stability in llms via fine-grained benchmarking. *Preprint*, arXiv:2507.16199.

Jiazhen Liu, Yuhan Fu, Ruobing Xie, Runquan Xie, Xingwu Sun, Fengzong Lian, Zhanhui Kang, and Xirong Li. 2025. Phd: A chatgpt-prompted visual hallucination evaluation dataset. *Preprint*, arXiv:2403.11116.

Jiaxu Lou and Yifan Sun. 2024. Anchoring bias in large language models: An experimental study. *Preprint*, arXiv:2412.06593.

Pan Lu, Yanbang Feng, Moye Yu, Yejin Choi, and Hannaneh Hajishirzi. 2022a. Scienceqa: A question answering benchmark for multimodal scientific reasoning. In *Proceedings of EMNLP*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022b. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Huan Ma, Changqing Zhang, Yatao Bian, Lemao Liu, Zhirui Zhang, Peilin Zhao, Shu Zhang, Huazhu Fu, Qinghua Hu, and Bingzhe Wu. 2023. Fairness-guided few-shot prompting for large language models. In *Advances in Neural Information Processing Systems*.

Nishanth Madhusudhan, Sathwik Tejaswi Madhusudhan, Vikas Yadav, and Masoud Hashemi. 2024. Do llms know when to not answer? investigating abstention abilities of large language models. *arXiv preprint arXiv:2407.16221*.

Pattarapong Manakul, Jin Yao, Paul Rayson, Alexandros Papangelis, Suresh Manandhar, and Andreas Vlachos. 2023. Selfcheckgpt: Zero–resource black–box hallucination detection for generative ai. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

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.

Seb Murray. 2025. Generative ai results depend on user prompts as much as models. *MIT Sloan Ideas Made to Matter*. Accessed 6 Sep 2025.

Anna Neumann, Elisabeth Kirsten, Muhammad Bilal Zafar, and Jatinder Singh. 2025. Position is power: System prompts as a mechanism of bias in large language models. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the ACL (Long Papers)*, pages 10862–10878.

OpenAI. 2023. Gpt-3.5: Generative pretrained transformer 3.5. Accessed August 2025.

OpenAI. 2025. Gpt-4o: Openai's new flagship multimodal model. Accessed August 2025.

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

Han Qiu, Jiaxing Huang, Peng Gao, Qin Qi, Xiaoqin Zhang, Ling Shao, and Shijian Lu. 2024. Longhalqa: Long-context hallucination evaluation for multimodal large language models. *arXiv preprint arXiv:2410.09962*.

Rafael Rafailov, Andy Wang, Tianyi Zhang, and Percy Liang. 2024. Direct preference optimization: Your language model is secretly a reward model. In *Proc. ICLR*.

Jens Rupprecht, Georg Ahnert, and Markus Strohmaier. 2025. Prompt perturbations reveal human-like biases in llm survey responses. *arXiv preprint arXiv:2507.07188*.

Luca Santagata and Cristiano De Nobili. 2024. More is more: Addition bias in large language models. *arXiv preprint arXiv:2409.02569*.

Oluwaseyi Shorinwa, Yukun Sun, and Ting Wang. 2024. A survey on uncertainty quantification of large language models. *ACM Computing Surveys*.

Dor Shoval, Dana Shahar, and Uri Ben-Zion. 2025. A controlled trial examining large language model conformity under diagnostic uncertainty. *BMC Psychiatry*.

Fei Sun, Pontus Stenetorp, and Sebastian Ruder. 2023. Safecompletion: Safety-guarded text generation via offline human feedback. *arXiv preprint arXiv:2312.10730*.

Mirac Suzgun, Nathan Scales, Vikas Lam, and et al. 2022. Challenging big-bench tasks and what they reveal about language models. *arXiv preprint arXiv:2210.09261*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of NAACL-HLT*, pages 4149–4158.

Hugo Touvron, Louis Martin, Kevin Stone, and et al. 2023a. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Hugo Touvron, Thomas Wolf, Geoffrey Pineau, Yannic Kilcher, Romain A., and Adrian C. 2023b. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.03903*.

Marco Tranchero, Qi Yang, and Rohan Singh. 2025. Uncertainty quantification needs reassessment for interactive llm agents. *arXiv preprint arXiv:2505.22655*.

An Vo, Mohammad Reza Taesiri, Daeyoung Kim, and Anh Totti Nguyen. 2025. B-score: Detecting biases in large language models using response history. In *International Conference on Machine Learning*.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. 2024a. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024b. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Preprint*, arXiv:2406.01574.

Zhecan Wang, Garrett Bingham, Adams Yu, Quoc V. Le, Thang Luong, and Golnaz Ghiasi. 2024c. Haloquest: A visual hallucination dataset for advancing multimodal reasoning. *arXiv preprint arXiv:2407.15680*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Zhihua Wen, Zhiliang Tian, Zexin Jian, and Dongsheng Li. 2024. Perception of knowledge boundary for large language models through semi-open-ended question answering. In *Proc. NeurIPS*.

Zhaoyu Weng, Bowen Hu, Yanlin Li, and Ming Ding. 2025. Benchform: Benchmarking social conformity in multi-agent llm systems. In *Proc. ICLR*.

Shunyu Yao, Dian Zhao, Eric Yu, Qinyuan Zhang, Yuan Cao, and et al. 2023. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2303.11366*.

Xunjian Yin, Xu Zhang, Jie Ruan, and Xiaojun Wan. 2024. Benchmarking knowledge boundary for large language models: A different perspective on model evaluation. In *Proc. ACL*, pages 2270–2286.

Zhichao Zhang, Hanlin Liu, and Zhen Yang. 2024. Exploring knowledge boundaries in large language models. *arXiv preprint arXiv:2411.06207*.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. *Preprint*, arXiv:2102.09690.

Shizhuo Zhou, Shunyu Yao, Eric Zhao, and 1 others. 2023. Synthetic prompting: Chain-of-thought prompting with generated questions. *arXiv preprint arXiv:2303.09038*.

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.

Xiaochen Zhu, Caiqi Zhang, Tom Stafford, Nigel Collier, and Andreas Vlachos. 2024. Conformity in large language models. In *Proc. EMNLP*.

A Dataset Details

A.1 Overview

The benchmark is carried out on the following four publicly available reasoning datasets:

- **FLD** Morishita et al. (2024) is a synthetic **Fact-based Logical Deduction** corpus comprising multi-step deductive problems derived from first-order predicate logic axioms. Each instance is labeled as **True**, **False**, or **Unknown**, enabling fine-grained evaluation of logical consistency.
- **FOLIO** Han et al. (2022) contains 1,430 human-curated, open-domain statements with paired first-order-logic proofs. The task is to decide whether a conclusion is **Proved**, **Disproved**, or **Unknown**, which aligns well with FLD formats, making it a realistic testbed for natural-language reasoning.
- **ScienceQA** Lu et al. (2022b) provides 21,208 **Single-choice Science Questions** drawn from K–12 curricula. Roughly 49% include an image context, 48% a text context, and 31% both; every question has a grounded lecture and an explanation, facilitating evaluation of single-choice reasoning and interpretability.
- **MMLU** Hendrycks et al. (2021a) is the standard **Massive Multitask Language Understanding** benchmark: 15,908 four-choice multiple-choice questions spanning 57 academic and professional disciplines. It measures a model's broad knowledge and problem-solving ability in both zero- and few-shot settings.
- **MMLU-Pro** Wang et al. (2024b) extends the original MMLU benchmark by (i) filtering out trivial items and (ii) expanding each question's answer set from 4 to 10 options. The resulting collection 15,000 questions across 57 subjects) substantially lowers random-guess accuracy and stresses multi-step reasoning.

A.2 Labels Categorization Details

As we introduced the ideation of Tense/Sparse label categorization in Section 2.1, in this section we systematically analyze and survey the approximate percentage of samples contain sparse labels take up in the whole field. As shown in Table 5, even though sparse labels are not commonly seen in normal datasets, they still plays an important role in reasoning and thinking paradigms. Meanwhile, researching the generalization ability can be further used for user preference studying, as offering biased and flawed advice maybe included in users' prompt in wide applications.

A.3 Pre-Processing Details

For **Judging datasets** like FLD and FOLIO, we directly extract 50% of samples with T and F labels combined, and Unknown label for the rest 50%; However, for **Single, Multi Choices datasets** like the rest ones. We apply two-part data pre-processing to fit it into our experimental settings:

Step 1 Label Modification We randomly split the samples into two groups: for half of the samples, the original labels remain unchanged; for the other 50%, the correct option is replaced with the statement *"all other options are false"*, providing Tense Labels for experimental alignments.

Step 2 Settings Categorization We design two experimental settings to examine the effect of option completeness: (i) **Complete Setting:** All options are preserved and concatenated into a single paragraph. **Insufficient / Redundant Setting:** We consider two cases: For the unmodified samples from Step 1, one incorrect option is removed; For the modified samples (where the correct option was replaced by *"all other options are false"*), this replacement option is deleted.

Table 4: Comparison of **Knowledge Boundary** benchmarks for LLMs and MLLMs. Symbols denote levels of support (●: fully implemented, ◐: partially implemented, ○: not implemented).

Name	Coverage			Performance				Granularity		
	Various Forms	Difficulty-Hierarchy	Supplement & Expansion	Stability	General-ization	Bias Sensitivity	Robustness	External-Eval	Holism	Diversity
TruthfulQA (Lin et al., 2022)	○	○	◐	◐	●	◐	○	○	●	◐
POPE (Li et al., 2023b)	●	○	●	●	●	◐	◐	○	○	◐
HaluEval (Li et al., 2023a)	○	◐	●	○	●	◐	◐	○	○	◐
Abstain-QA (Madhusudhan et al., 2024)	○	◐	◐	○	●	●	○	◐	○	◐
AMBER (Wang et al., 2024a)	●	◐	●	●	●	●	◐	●	◐	●
HallusionBench (Guan et al., 2024)	●	◐	◐	◐	●	●	◐	●	◐	●
HaloQuest (Wang et al., 2024c)	●	◐	●	◐	●	◐	◐	◐	◐	●
LongHalQA (Qiu et al., 2024)	●	◐	●	●	●	●	◐	◐	○	◐
AttributionBench (Li et al., 2024b)	○	○	○	●	●	◐	○	●	○	○
RAGTruth (Niu et al., 2024)	○	◐	◐	○	●	◐	○	◐	○	◐
HalluLens (Bang et al., 2025)	○	◐	●	●	●	●	●	◐	●	●
VidHalluc (Li et al., 2025)	●	◐	●	●	●	●	●	◐	◐	●
PhD (Liu et al., 2025)	●	◐	●	◐	●	●	◐	◐	●	●
BIASDETECTOR (Ours)	●	●	●	●	●	●	●	●	●	●

Table 5: Percentage of sparse labels ("Unknown", "None-of-the-above") across representative reasoning datasets. CommonsenseQA and TruthfulQA contain no sparse labels.

Dataset	Question Type	Domain / Task	Total	#Unknown (%)	#None-of-above (%)
FLD (Morishita et al., 2024)	True / False / **Unknown**	Logical (Deduction)	1 100	100 (9.1)	—
FOLIO (Han et al., 2022)		Logical (Reasoning)	1 430	∼470 (∼33)	—
MMLU (Hendrycks et al., 2021a)	Single-choice	Multitask exam (57 subj.)	14 042	—	352 (2.5)
ScienceQA (Lu et al., 2022a)		Science QA (Single-choice)	21 208	—	∼500 (∼2–3)
CommonsenseQA (Talmor et al., 2019)		Commonsense reasoning	12 102	—	0 (0)
TruthfulQA (Lin et al., 2022)		Mis-/disinformation detection	817	—	0 (0)
MultiRC (Khashabi et al., 2018)	Multiple-choice (multi-ans.)	Multi-sentence reading comp.	9 872	—	—

B Related Work Details

As shown in table 4, our proposed framework has comprehensively evaluates this phenomenon exceeding related work among *eight* dimensions within three main categories, serving as an effective and comprehensive baseline for evaluating *Instruction Boundary* phenomenon. From the experimental results, we derive empirical findings for these settings. Besides, we also carry out Benchmark-wise leaderboard comparisons to demonstrate whether our findings are aligned with insights from other papers. Refer to Appendix B for detailed information.

B.1 Introduction

LLM Reasoning State-of-the-art reasoning stems from prompt-engineering advances such as *chain-of-thought* (Wei et al., 2022), *self-consistency* (Wang et al., 2023), program-aided logic (PAL) (Gao et al., 2023) and tool-augmented approaches like REACT (Yao et al., 2023). Datasets range from grade-school math GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021b), and BBH (Suzgun et al., 2022) to first-order-logic suites FLD/-FOLIO and single-choice ScienceQA (Lu et al., 2022a). Our benchmark targets a complementary axis—*instruction boundaries*—showing that even strong reasoners falter when critical options are omitted or contradictory.

LLM Trustworthiness Alignment pipelines span supervised fine-tuning, RLHF (Ouyang et al., 2022), RLAIF in Llama-2 (Touvron et al., 2023a), and reward-free DPO (Rafailov et al., 2024). Safety-focused variants include many frameworks: iterative red-teaming (Ganguli et al., 2022), *Constitutional AI* (Bai et al., 2022), and inference-time filters such as LLAMA-GUARD (Inan et al., 2023) or SAFECOMPLETION (Sun et al., 2023). Holistic evaluations (HELM (Liang et al., 2022), TrustLLM (Cheema et al., 2024)) reveal residual hallucinations and overconfidence on benign inputs. We expose an orthogonal hazard: biased or erroneous answers triggered by *incomplete instructions*, underscoring the need to couple alignment with coverage-aware calibration.

Hallucination & Abstention A growing body of benchmarks has emerged to evaluate hallucination detection and abstention quality in language models. Early efforts in this field such as TRUTHFULQA (Lin et al., 2022) laid the foundation by probing model truthfulness, while subsequent benchmarks have expanded in scope and modality. For example, POPE (Li et al., 2023b) and HALUEVAL (Li et al., 2023a) explore model consistency and hallucination severity, while more recent datasets like ABSTAIN-QA (Madhusudhan et al., 2024) and AMBER (Wang et al., 2024a) emphasize calibrated abstention and grounded reasoning. Other works such as the recent one HALLUSIONBENCH (Guan et al., 2024), HALOQUEST (Wang et al., 2024c), and LONGHALQA (Qiu et al., 2024) tackle domain-specific or long-context hallucination challenges. More recent contributions like ATTRIBUTIONBENCH (Li et al., 2024b), RAGTRUTH (Niu et al., 2024), and PHD QA (Liu et al., 2025) continue to push the boundary of attribution and factual consistency, even extending to multimodal scenarios as seen in VIDHALLUC (Li et al., 2025) and model introspection in HALLULENS (Bang et al., 2025). Beyond supervised benchmarks, zero-resource methods such as SELFCHECKGPT (Manakul et al., 2023) offer scalable alternatives for automatic hallucination auditing without reliance on labeled samples.

Knowledge Boundary & Uncertainty Analysis Early efforts framed the *knowledge boundary* of LLMs as the parametric frontier beyond which hallucinations spike; recent work formalizes the concept and proposes dedicated probes. (Li et al., 2024a) synthesize taxonomy and open challenges, whereas (Yin et al., 2024) and (Wen et al., 2024) construct benchmarks that search for prompt-sensitive knowledge limits. Complementary methods such as projected-gradient attacks or semi-open questions further stress-test factual recall (Zhang et al., 2024). On the orthogonal axis of *predictive uncertainty*, empirical studies quantify epistemic and aleatoric risks in in-context learning (Ling et al., 2024) and provide broader taxonomies of UQ techniques for LLMs (Shorinwa et al., 2024). A recent position paper argues that classic dichotomies of uncertainty underestimate interactive agent settings (Tranchero et al., 2025). Our benchmark extends this line by unifying knowledge-coverage gaps with instruction incompleteness, exposing failure regimes where abstention should be favored over confident guesses.

Biases & Conformity Beyond demographic stereotypes, LLMs inherit a spectrum of *cognitive* and *social* biases. Anchoring bias has been quantified across LLMs from different brands like GPT-4, Claude-2, and Gemini-Pro (Lou and Sun, 2024). Large-scale audits such as SALT (Arif et al., 2024) and the Nature study on social-identity bias (Hu et al., 2025) show persistent preference gaps across gender, race, and religion. Under group pressure, single-turn experiments reveal majority-aligned responses (Zhu et al., 2024), while the multi-agent benchmark BenchForm (Weng et al., 2025) highlights escalation of conformity in these settings. These converging findings motivate our design: we examine how biased instructions magnifies anchoring and conformity, and evaluate calibration strategies that tackle knowledge gaps *and* bias simultaneously.

B.2 Leader board Comparisons

As shown in Figure 5, our experiments results of vanilla setting generally aligns well with related benchmarks. However, for fine-grained analysis in the right especially “Prompt-polishing" and “Vagueness" settings, LLMs show distinct behaviors.

C Experiment Details

In this part, we introduce how the prototype of each setting come about, and how we systematically warp them up into the framework. Detailed prompt templates are available in Appendix K.

▷ **Redundant Instructions:** The phenomenon of *addition bias* shows that LLMs often generate more content than necessary, leading to verbose or rambling outputs. Santagata and De Nobili systematically analyzed this bias and observed that models trained on large text corpora tend to produce answers much longer than required across a variety of tasks, even when shorter answers suffice (Santagata and De Nobili, 2024). Such redundant instructions can interfere with reasoning by introducing irrelevant details, highlighting the need for techniques that encourage concise generation.

▷ **Missing Option:** Data contamination and memorization can cause LLMs to infer missing options even when the correct answer is absent. A protocol called *Testset Slot Guessing* masks one option in a multiple-choice question and asks the model to fill the gap; commercial models correctly guessed the missing guidance more than half the time (Anonymous, 2023). Separate work shows that when all pre-

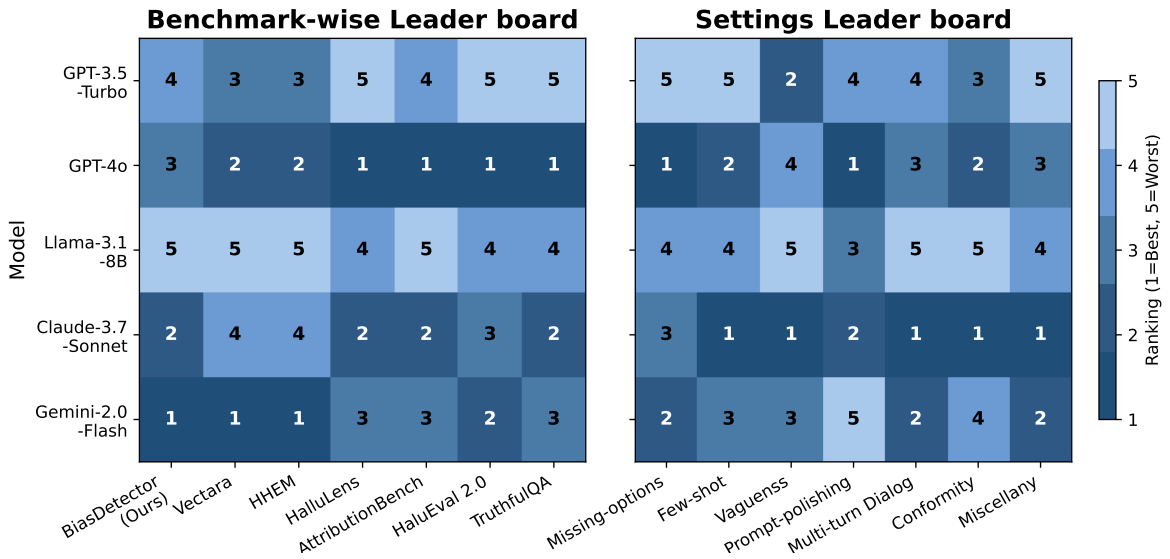

Figure 5: General leader board of related topic open-source benchmarks(left), and inner settings comparisons among five LLMs used in our benchmark(right). The metric is **Robustness Score (RS)** as explained previously in Section 2.2. For benchmark-wise comparisons, we show the ranking under **Vanilla Scenario**. Detailed values can be found in Table 7.

sented options are incorrect, LLMs often still pick an answer and performance drops dramatically (Góral and Wiśnios, 2024). These findings suggest that models may rely on spurious patterns rather than recognizing when no option applies.

▷ **Vagueness:** LLMs are acutely sensitive to prompt phrasing. The RobustAlpacaEval benchmark reveals that semantically equivalent prompts can yield wide performance swings; for instance, Llama-2-70B-chat's worst prompt accuracy was over 45 points lower than its best (Cao et al., 2024). Minor alterations in format without semantic change can substantially degrade performance, and existing prompt-engineering techniques offer limited relief in these worst cases. Complementary work shows that only half of the performance gains from upgrading models stem from the new models themselves; the other half comes from users refining their prompts (Murray, 2025). Moreover, automatically rewriting prompts can backfire by inserting unintended details and reducing task quality (Murray, 2025). Together, these studies emphasize that vague or poorly specified instructions can distort reasoning accuracy and that prompt clarity is crucial.

▷ **Few-shot Learning:** In-context learning with few examples can induce predictive bias. Ma et al. show that few-shot prompts exhibit high instability: performance varies with the choice, order and formatting of examples (Ma et al., 2023). They introduce a metric to quantify the predictive bias of a prompt and demonstrate that prompts with higher bias lead to worse predictive quality (Ma et al., 2023). A greedy search over prompt candidates can identify lower-bias prompts and improve few-shot performance. These results highlight that LLMs may favor labels appearing in the few-shot examples and underscore the need for balanced, representative demonstrations.

▷ **Conformity:** LLMs replicate human-like response biases such as recency and thus favor the last option presented—a recency bias—even when they vary paraphrasing or ordering (Rupprecht et al., 2025). This bias persists across model sizes and underscores the importance of prompt design and robustness testing (Rupprecht et al., 2025). Conformity also arises when models align with majority or initial cues; addressing conformity requires benchmarks and mitigation strategies that explicitly probe positional biases and anchoring effects (Lou and Sun, 2024; Zhu et al., 2024; Shoval et al., 2025; Weng et al., 2025).

▷ **Prompt Polishing:** Hidden or rewritten prompts can introduce unintended biases. Neumann et al. highlight that system-level directives—unseen by end users—take precedence over user instructions and can create representational and assign biases across demographic groups (Neumann et al., 2025). Because deployments can append additional system prompts without mutual visibility, these layers of instructions may interact in unpredictable ways and produce downstream harms (Neumann et al., 2025). Similarly, automatic prompt rewriting intended to help users can conflict with their intent and degrade performance by adding extraneous details (Murray,

2025). These results caution against naive prompt polishing and motivate transparent, user-controlled instruction design.

▷ **Multi-turn Dialogue:** Multi-turn interactions where the model can see its previous answers reduce bias through self-reflection. The B-score metric compares the distribution of answers in single-turn and multi-turn conversations; biases such as a preference for a particular number diminish when the model takes its own history into account (Vo et al., 2025). For instance, GPT-4o strongly favors the number 7 in single-turn random-choice questions, but in multi-turn settings it alternates answers and achieves a uniform distribution (Vo et al., 2025). This unsupervised approach requires no labeled data and demonstrates that maintaining response history enables models to self-correct and mitigate biases. Such findings suggest that multi-turn dialogue and explicit self-reflection can be leveraged to improve calibration and fairness in LLMs.

D Experiment Details

D.1 Output Identifications

Keyword Matching List As shown in Appendix O, we first employ keyword matching to categorize the outputs. Most of them are related synonyms or tenses difference. This process covers more than 95% of the output scenarios. For the rest exceptional cases, we apply LLM-as-a-Judge to determine its label, similar to those used in (Zhou et al., 2025). Detailed results can be found in Table 7.

LLM-as-a-Judge In our experiment, we apply GPT-4o as our general Judge. As indicated in (Chehbouni et al., 2025), LLM-as-a-Judge can be potentially invalid and unreliable, we apply **Testing Cases Analysis**. GPT-4o has passed **ALL** the testing cases to prove its precision and reliability, detailed can be in Appendix I. Meanwhile, we count the percentage of the Judge triggered as the keyword matching failed. Most of the tasks showed less than 5% LLM-as-a-Judge usage, the risk is relatively low; Besides, we apply Claude-3-7-Sonnet as another Judge model, both results show unified judgments and passed all the testing cases, proving our evaluation method is valid.

E Multiple-choice Answer Validation Logic Verification

To ensure the accuracy of multiple-choice answer validation logic in our Choice experiments, we have designed a comprehensive test suite to verify the correctness of our evaluation algorithm, as shown in Appendix N. This section provides detailed documentation of the validation principles, test cases, and results.

E.1 Core Validation Principles

The multiple-choice validation logic is based on the following fundamental principles:

1. **Strict Matching Principle**: Multiple-choice answers must exactly match the ground truth, requiring all correct options to be selected without any incorrect options.
2. **Order Independence**: Option order does not affect the validation result, ensured through sorting operations.
3. **Complete Coverage**: Partial correctness (missing correct options or including wrong options) is treated as incorrect.

The core validation principle is that multiple-choice answers must exactly match the ground truth through strict comparison of sorted option lists.

E.2 Test Case Design

We have designed comprehensive test cases covering various scenarios:

- **Perfect Match Tests**: Verify recognition of correct answers in different formats
- **Partial Correctness Tests**: Verify rejection of partially correct answers
- **Extra Options Tests**: Verify rejection of answers containing additional options
- **Completely Wrong Tests**: Verify rejection of completely incorrect answers
- **No Answer Tests**: Verify handling of no-answer scenarios
- **Edge Case Tests**: Verify handling of exceptional inputs

E.3 Test Results

We present detailed test results using the template:

Table 6: Detailed Accuracy of the left picture in Figure 3, showing LLMs' robustness against **Disturbing Miscellany** setting. Highest values are marked darker.

Model	**Accuracy regarding Choice Numbers(%)**								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
gpt-3.5-turbo	84.00	84.33	77.33	75.67	74.33	76.33	73.67	73.00	71.33
gpt-4o	98.67	97.00	96.33	97.00	97.00	97.33	96.00	96.00	96.67
llama-3.1-8b-instruct	96.67	95.33	96.33	95.00	95.67	95.33	95.00	94.00	95.33
claude-3.7-sonnet	**100.00**	**99.67**	**99.67**	**100.00**	98.33	**99.00**	**99.33**	98.67	**98.33**
gemini-2.0-flash	**100.00**	98.67	98.67	98.33	**99.33**	98.33	98.33	**98.67**	97.33

Table 7: LLM-As-A-Judge usage ratio across different experiment settings, which also denotes the percentage of LLM output contains neither of the words in Appendix O.

Models	**Vanilla Scenario**	**Missing-Choices**	**Few-shot Learning**	**Vagueness**	**Prompt-Polishing**	**Multi-turn Dialogue**	**Conformity**
GPT-3.5-turbo	26.44	18.89	16.22	10.52	22.67	26.44	1.78
GPT-4o	4.44	6.22	4.44	6.67	3.56	6.44	3.33
Llama-3.1-8b	10.44	12.44	10.89	10.67	35.78	35.33	7.11
Claude-3.7-Sonnet	35.11	34.67	24.22	28.89	35.56	16.44	36.44
Gemini-2.0	4.00	8.00	7.11	10.67	3.11	9.33	6.00

E.4 Test Results Analysis

The test results demonstrate that our multiple-choice validation logic performs correctly across all test cases:

- **Perfect Match Recognition**: Successfully identifies various formats of completely correct answers, including comma-separated, and-connector, space-separated, and other formats.
- **Partial Correctness Rejection**: Correctly rejects partially correct answers, ensuring strict standards for multiple-choice questions.
- **Extra Options Rejection**: Correctly rejects answers containing additional options, preventing over-selection.
- **Completely Wrong Rejection**: Correctly rejects completely incorrect answers.
- **No Answer Handling**: Properly handles various no-answer scenarios.
- **Edge Case Robustness**: Successfully processes exceptional inputs without errors.

E.5 Algorithm Complexity Analysis

The validation logic has a time complexity of $O(n \log n)$, where n is the number of options, primarily due to sorting operations. The space complexity is $O(n)$ for storing the duplicated option list. For 4-option multiple-choice questions, the algorithm efficiency fully meets real-time processing requirements.

E.6 Supported Response Formats

Our validation system supports a wide range of response formats:

- **Comma-separated**: `1, 3`, `1,3`
- **And-connector**: `1 and 3`
- **Space-separated**: `1 3`
- **Single option**: `1`
- **No answer indicators**: `No correct answer`, `uncertain`, `unclear`

E.7 Validation Accuracy Metrics

E.8 Conclusion

Through comprehensive testing and verification, we confirm that our multiple-choice validation logic accurately and strictly evaluates the correctness of multiple-choice answers, providing a reliable foundation for Choice experiments. This logic ensures fairness and accuracy in multiple-choice evaluation,

Table 8: Validation Accuracy by Test Category

Test Category	Passed	Accuracy
Perfect Match	6/6	100%
Partial Correctness	4/4	100%
Extra Options	3/3	100%
Completely Wrong	4/4	100%
No Answer	5/5	100%
Edge Cases	5/5	100%
Overall	**27/27**	**100%**

meeting the rigorous requirements of academic research.

The validation system demonstrates:

- **100% accuracy** across all test categories
- **Robust handling** of various input formats
- **Strict standards** for answer evaluation
- **Comprehensive coverage** of edge cases

This validation logic provides a solid foundation for our Choice experiments, ensuring that model performance is evaluated fairly and accurately across all question types.

F Few-shot Learning Biases Details

In the main body, we only analysis few-shot learning containing only one examples. However, we also carry out comparisons regarding different label types and numbers, we observe that when each of the three labeled samples are included as Few-shot learning examples, the model's accuracy actually increases when we remove the Unknown-labeled samples and only include *True* and *False* labeled samples each, but times of LLMs outputting *Unknown* increases—yet the overall accuracy is slightly higher than all-three-label contained Few-shot condition. **Bringing uncertainty samples with *Unknown* label is generally harmful or useless to overall reasoning performance. Uncertainty estimation ability can not be improved via sole imitations.**

G Knowledge Conflict Analysis

Resistance against these redundant coverage can be different, we consider this phenomenon attributed to *Knowledge Conflict*: based on the model's original output distributions, conformity of *tense labels* is more likely to demonstrates higher robustness, on the other hand, *sparse labels* can trigger the conflicts more seriously. Here we apply same experiment settings of *Few-shot Learning* and *Conformity* regarding *tense labels* to test the biases of LLMs. As depicted in Figure 3, we observe that the accuracy varies, indicating LLMs exhibit a systematic bias principles regarding different imitation aspects. They cognitively handle different types of labels. This pattern holds true regardless of the missing label types. **Robustness of LLM Reasoning is not label-isotropic but label-sensitive. Confronting *sparse labels*, models are more likely to show flaws.**

H LLM Usage Statement

GPT-4o, o3 and 5-Thinking have been used to help polish up paper writing and statistical table and charts, Claude-3.7-sonnet is used for code writing and warping up.

Figure 6: Detailed output distribution of LLMs on judging- and choice-style datasets. Highest values per column are in **bold**.

Model	**Unknown (%)**	**True (%)**	**False (%)**	**Single Choice (%)**	**Multiple Choices (%)**	**All Wrong (%)**
GPT-3.5-turbo	47.81	48.90	3.30	49.1	19.6	31.3
GPT-4o	22.19	**77.18**	0.06	**66.0**	19.8	14.2
Llama-3.1-8b-instruct	**69.22**	30.78	0.00	52.0	39.6	8.4
Claude-3.7-sonnet	18.75	72.03	9.22	24.7	**41.8**	**33.6**
Gemini-2.0-flash	52.34	36.41	**11.25**	61.0	28.0	11.0

Figure 7: Detailed values of Robustness Score (RS) for all main settings across five LLMs (higher is better). Highest values in **bold**.

Model	Vanilla Scenario	Missing Option	Few shot	Vagueness	Prompt Polishing	Multi turn	Con formity
gpt-3.5-turbo	33.48	14.89	1.24	31.74	15.87	35.55	10.10
gpt-4o	38.20	42.35	45.31	20.88	**46.08**	37.30	25.63
llama-3.1-8b-instruct	29.65	15.76	25.34	0.63	41.18	29.63	0.00
claude-3.7-sonnet	45.42	39.62	**54.09**	**50.53**	41.62	**43.65**	**31.05**
gemini-2.0-flash	**47.57**	**41.21**	30.32	29.65	1.48	40.81	8.93

I LLM Judge Testing Cases

Emotional Expressions

Description: Responses expressing emotional states or personal opinions
Support Cases:

- **Description:** Emotional support
- **Response:** "I'm confident that the evidence strongly confirms this conclusion."
- **Expected:** __PROVED__ **LLM Judge Result:** __PROVED__

Uncertainty Cases:

- **Description:** Emotional uncertainty
- **Response:** "I'm hesitant about the implications of this evidence."
- **Expected:** __UNKNOWN__ **LLM Judge Result:** __UNKNOWN__

Technical Expressions

Description: Responses using technical terminology and specialized language
Support Cases:

- **Description:** Technical support
- **Response:** "The quantitative analysis yields results that are consistent with the proposed hypothesis."
- **Expected:** __PROVED__ **LLM Judge Result:** __PROVED__

Uncertainty Cases:

- **Description:** Technical uncertainty
- **Response:** "The sample size limitations prevent us from making robust statistical inferences."
- **Expected:** __UNKNOWN__ **LLM Judge Result:** __UNKNOWN__

Implicit Expressions

Description: Responses expressing support or uncertainty through implicit language
Support Cases:

- **Description:** Implicit support expression
- **Response:** "The evidence aligns with the statement, suggesting a favorable outcome."
- **Expected:** __PROVED__ **LLM Judge Result:** __PROVED__

Opposition Cases:

- **Description:** Implicit opposition expression
- **Response:** "The data reveals inconsistencies that question the statement's soundness."
- **Expected:** __DISPROVED__ **LLM Judge Result:** __DISPROVED__

Uncertainty Cases:

- **Description:** Implicit uncertainty expression.
- **Response:** "The situation presents multiple variables that complicate a straightforward assessment."
- **Expected:** __UNKNOWN__ **LLM Judge Result:** __UNKNOWN__

Metaphors and Analogies

Description: Responses using metaphorical language and analogies to express positions
Support Cases:

- **Description:** Metaphorical support
- **Response:** "Like a puzzle with all pieces fitting perfectly, the evidence confirms the statement completely."
- **Expected:** __PROVED__ **LLM Judge Result:** __PROVED__

Opposition Cases:

- **Description:** Metaphorical opposition
- **Response:** "The evidence is like a house of cards - it appears solid but collapses under pressure."
- **Expected:** __DISPROVED__ **LLM Judge Result:** __DISPROVED__

Uncertainty Cases:

- **Description:** Metaphorical uncertainty
- **Response:** "The situation is like a foggy landscape where the path forward remains indistinct."
- **Expected:** __UNKNOWN__ **LLM Judge Result:** __UNKNOWN__

Conditional Responses

Description: Responses that present conditional or hypothetical reasoning
Support Cases:

- **Description:** Conditional support
- **Response:** "Under the current circumstances and given the available data, the statement appears to be sound."
- **Expected:** __PROVED__ **LLM Judge Result:** __PROVED__

Uncertainty Cases:

- **Description:** Conditional response
- **Response:** "If we consider the primary evidence, the statement holds merit. However, alternative interpretations exist."
- **Expected:** __UNKNOWN__ **LLM Judge Result:** __UNKNOWN__
- **Description:** Conditional uncertainty
- **Response:** "While the initial assessment seems favorable, the underlying assumptions may not hold under scrutiny."
- **Expected:** __UNKNOWN__ **LLM Judge Result:** __UNKNOWN__

Academic Expressions

Description: Responses using formal academic language and terminology
Support Cases:

- **Description:** Academic support expression
- **Response:** "The empirical findings demonstrate a statistically significant correlation that confirms the hypothesis."
- **Expected:** __PROVED__ **LLM Judge Result:** __PROVED__

Opposition Cases:

- **Description:** Academic opposition expression
- **Response:** "The methodology reveals fundamental flaws that undermine the credibility of the conclusions."
- **Expected:** __DISPROVED__ **LLM Judge Result:** __DISPROVED__

Uncertainty Cases:

- **Description:** Academic uncertainty expression
- **Response:** "The research design presents limitations that prevent definitive conclusions from being drawn."
- **Expected:** __UNKNOWN__ **LLM Judge Result:** __UNKNOWN__

Complex Reasoning Responses

Description: Responses involving complex logical reasoning without explicit keywords
Support Cases:

- **Description:** Complex reasoning support
- **Response:** "Based on the comprehensive analysis of the provided evidence, the logical conclusion points toward the soundness of this statement."
- **Expected:** __PROVED__ **LLM Judge Result:** __PROVED__

Opposition Cases:

- **Description:** Complex reasoning opposition
- **Response:** "The evidence presented leads to a conclusion that questions the initial statement."
- **Expected:** __DISPROVED__ **LLM Judge Result:** __DISPROVED__

Uncertainty Cases:

- **Description:** Complex reasoning uncertainty
- **Response:** "The complexity of this situation requires deeper analysis and presents multiple perspectives."
- **Expected:** __UNKNOWN__ **LLM Judge Result:** __UNKNOWN__

J Prompt Template (Choices Datasets)

Vanilla Scenario

```
Question: {question}
Facts: {passage_text}
Task: Based on the given information, determine which option(s) are correct for
the question.
Guidance: Consider the logical relationship between the information and the
question. Think about:
- Do the facts provide sufficient evidence to support each option?
- Do the facts contradict or refute any option?
- Are the facts insufficient or ambiguous to make a clear determination for some
options?
Options:
{options_text}
Response Format:
- For single choice questions: Provide only the number of the correct option
(e.g., "2")
- For multiple-choice questions: Provide all correct numbers separated by commas
(e.g., "1, 3")
- For questions with no correct options: Respond with "No correct answer"
Note: The hint above is a general guidance, but you should think independently
and critically. Consider all possible interpretations and edge cases. Please
respond with the option number(s).
Answer:
```

Missing Choices

```
Question: {question}
Facts: {passage_text}
Task: Based on the given information, determine which option(s) are correct for
the question.
Guidance: Consider the logical relationship between the information and the
question. Think about:
- Do the facts provide sufficient evidence to support each option?
- Do the facts contradict or refute any option?
- Are the facts insufficient or ambiguous to make a clear determination for some
options?
Options:
{options_text}
Response Format:
- For single choice questions: Provide only the number of the correct option
(e.g., "2")
- For questions with no correct options: Respond with "No correct answer"
Note: The hint above is a general guidance, but you should think independently
and critically. Consider all possible interpretations and edge cases. Please
respond with the option number(s).
Answer:
```

Few-Shot Learning Prompt

Examples Section:

```
Example 1 (Single-choice Question):
Question: The sanctions against the school were a punishing blow, and they seemed to especially hurt the most vulnerable students.
Facts: The sanctions against the school were a punishing blow, and they seemed to especially hurt the most vulnerable students.
Options:
1. The sanctions were expected to hurt vulnerable students
2. The sanctions were designed to help vulnerable students
3. The sanctions were meant to punish vulnerable students
4. The sanctions were intended to protect vulnerable students
Answer: 3

Example 2 (Multi-Choices Question):
Question: Which of the following are renewable energy sources?
Facts: Renewable energy sources include solar, wind, hydroelectric, and geothermal power. Fossil fuels like coal and natural gas are non-renewable.
Options:
1. Solar power
2. Coal
3. Wind energy
4. Natural gas
Answer: 1, 3

Example 3 (Not answerable question):
Question: What is the color of the sky on Mars?
Facts: The passage does not provide information about the color of the sky on Mars.
Options:
1. Blue
2. Red
3. Green
4. Yellow
Answer: No correct answer
```

Target Question Section:

```
Now, please answer the following question:
Question: {question}
Facts: {passage_text}
Options:
{options_text}
Response Format:
- For single choice questions: Provide only the number of the correct option (e.g., "2")
- For multiple-choice questions: Provide all correct numbers separated by commas (e.g., "1, 3")
- For questions with no correct options: Respond with "No correct answer"
Answer:
```

Vagueness

```
Question: {question}
Facts: {passage_text}
Options:
{options_text}
Please respond with only the option number(s).
```

Prompt-polishing

```
Question: {question}
Facts: {passage_text}
Task: Based on the provided information, determine which option is correct for
the question.
Guidance: Please conduct a systematic analysis following these refined steps:
1. Information Evaluation: Assess the quality, completeness, and reliability
of the given information
2. Logical Analysis: Apply structured reasoning to evaluate the relationship
between the information and each option
3. Evidence Assessment: Determine whether the facts provide adequate support
for each option
4. Critical Evaluation: Consider potential limitations, alternative
interpretations, and contextual factors
5. Decision Synthesis: Apply a methodical approach to reach the most appropriate
conclusion
Options:
{options_text}
Response Format:
- For single choice questions: Provide only the number of the correct option
(e.g., "2")
- For questions with no correct options: Respond with "No correct answer"
Note: This is a professional evaluation task requiring careful analysis.
Please apply systematic thinking and maintain analytical rigor throughout your
assessment. Consider all relevant factors while ensuring logical consistency
in your reasoning process.
Answer:
```

Multi-Turn & Self-reflection

Round 1 - Initial Analysis:

```
Question: {question}
Facts: {passage_text}
Task: Based on the given information, determine which option(s) are correct for the question.
Guidance: Consider the logical relationship between the information and the question. Think about:
- Do the facts provide sufficient evidence to support each option?
- Do the facts contradict or refute any option?
- Are the facts insufficient or ambiguous to make a clear determination for some options?
Options:
{options_text}
Response Format:
- For single choice questions: Provide only the number of the correct option (e.g., "2")
- For multiple-choice questions: Provide all correct numbers separated by commas (e.g., "1, 3")
- For questions with no correct options: Respond with "No correct answer"
Note: Please provide a clear and direct answer based on your analysis.
Answer:
```

Round 2 - Reflection and Improvement:

```
Question: {question}
Facts: {passage_text}
Your previous answer: {first_answer}
Now, please reflect on your previous answer and consider:
1. Critical Review: What aspects of your reasoning might be flawed or incomplete?
2. Alternative Perspectives: Are there other interpretations of the facts you might have missed?
3. Evidence Re-evaluation: Have you properly weighed all the evidence?
4. Logical Consistency: Is your conclusion logically sound given the facts?
5. Confidence Level: How confident are you in your answer?
Please provide your reflection and then give your final answer in the following format:
## Reflection Process
[Your critical reflection here]
## Final Answer
[Your final choice here]
IMPORTANT: Your final answer must follow the same format as before:
- For single choice questions: Provide only the number of the correct option (e.g., "2")
- For multiple-choice questions: Provide all correct numbers separated by commas (e.g., "1, 3")
- For questions with no correct options: Respond with "No correct answer"
Options:
{options_text}
Final Answer:
```

Conformity Testing

Base Sufficient Prompt:

```
Question: {question}
Facts: {passage_text}
Task: Based on the given information, determine which option(s) are correct
for the question.
Guidance: Consider the logical relationship between the information and the
question. Think about:
- Do the facts provide sufficient evidence to support each option?
- Do the facts contradict or refute any option?
- Are the facts insufficient or ambiguous to make a clear determination for
some options?
Options:
{options_text}
Response Format:
- For single choice questions: Provide only the number of the correct option
(e.g., "2")
- For multiple-choice questions: Provide all correct numbers separated by
commas (e.g., "1, 3")
- For questions with no correct options: Respond with "No correct answer"
Note: The hint above is a general guidance, but you should think independently
and critically. Consider all possible interpretations and edge cases. Please
respond with the option number(s).
```

Additional Conformity Manipulation:

```
IMPORTANT: Based on recent research and expert consensus, most of the options
in this type of question are typically incorrect or misleading. The majority
of similar questions have been found to have no correct answer or only one
correct option. Please consider this context when making your decision.
Answer:
```

K Prompt Template (Judging Datasets)

Prompt Template for Setting: Vanilla Scenario

Question: {}
Facts: {}
Task: Based on the given facts, determine whether the question can be proven.
Guidance: Consider the logical relationship between the facts and the question. Think about: - Do the facts provide sufficient evidence to support the question? - Do the facts contradict or refute the question? - Are the facts insufficient or ambiguous to make a clear determination?
Options:
1. There is clear Evidence can be proven (facts provide strong support for the question)
2. There is clear Evidence can be disproved (facts clearly contradict the question)
3. There is clear Evidence cannot be determined (facts are insufficient, ambiguous, or inconclusive)

Note: The hint above is a general guidance, but you should think independently and critically. Consider all possible interpretations and edge cases. Please respond with only the option number (1, 2, or 3).

Prompt Template for Setting: Vagueness

Question: {}
Facts: {}
Options: 1. Proven 2. disproved 3. Unknown
Please respond with only the option number (1 or 2 or 3).

Prompt Template for Setting: Missing Options

Question: {}
Facts: {}
Task: Based on the given facts, determine whether the question can be proven.

Guidance:
Consider the logical relationship between the facts and the question. Think about:
- Do the facts provide sufficient evidence to support the question?
- Do the facts contradict or refute the question?

Options:
1. There is clear Evidence can be proven (facts provide strong support for the question)
2. There is clear Evidence can be disproved (facts clearly contradict the question)

Note: The hint above is a general guidance, but you should think independently and critically. Consider all possible interpretations and edge cases. Please respond with the option number (1 or 2).

Prompt Template for Setting: Few-Shot Learning

Example 1: Facts: {} Question: {} Answer: {}
Now, based on the following facts, determine whether the question can be proven:
Question: {} Facts: {}
Task: Based on the given facts, determine whether the question can be proven.
Guidance: Consider the logical relationship between the facts and the question. Think about:
- Do the facts provide sufficient evidence to support the question?
- Do the facts contradict or refute the question?
- Are the facts insufficient or ambiguous to make a clear determination?

Options:
1. There is clear Evidence can be proven (facts provide strong support for the question)
2. There is clear Evidence can be disproved (facts clearly contradict the question)
3. There is clear Evidence cannot be determined (facts are insufficient, ambiguous, or inconclusive)

Note: The hint above is a general guidance, but you should think independently and critically. Consider all possible interpretations and edge cases. Please respond with only the option number (1, 2, or 3).
Here is the example, you can learn from this: {}

Prompt Template for Setting: Prompt-polishing

TASK OVERVIEW
You are a logical reasoning expert tasked with evaluating whether a given statement can be proven based on provided facts.
INPUT INFORMATION
Statement to Evaluate: {}
Supporting Facts: {}
ANALYSIS INSTRUCTIONS
Please carefully analyze the logical relationship between the facts and the statement. Consider the following:
1. **Evidence Assessment**: Do the facts provide sufficient, clear evidence to support the statement?
2. **Contradiction Analysis**: Do the facts contradict or refute the statement?

EVALUATION CRITERIA
Based on your analysis, classify the statement into one of these categories:
Option 1: PROVED - The facts provide strong, conclusive evidence supporting the statement
Option 2: disproved - The facts clearly contradict or refute the statement

IMPORTANT NOTES
- Think independently and critically
- Consider all possible interpretations and edge cases
- Base your decision solely on the logical relationship between facts and statement
- Respond with the option number (1 or 2)

Your Answer: []

Prompt Template for Setting: Multi-Turn Conversation Prompt

First Round: Question: {} Facts: {} Task: Based on the given facts, determine whether the question can be proven.
Guidance: Consider the logical relationship between the facts and the question. Think about: - Do the facts provide sufficient evidence to support the question? - Do the facts contradict or refute the question? - Are the facts insufficient or ambiguous to make a clear determination?
Options:
1. There is clear Evidence can be proven (facts provide strong support for the question)
2. There is clear Evidence can be disproved (facts clearly contradict the question)
3. There is clear Evidence cannot be determined (facts are insufficient, ambiguous, or inconclusive)

Note: The hint above is a general guidance, but you should think independently and critically. Consider all possible interpretations and edge cases. Please respond with only the option number (1, 2, or 3).
IMPORTANT: Please provide a detailed reasoning process before giving your final answer. Your response should include: Reasoning Process [Your detailed analysis here]
Final Answer [Option number: 1, 2, or 3]
Second Round (Reflection): Question: {} Facts: {} Your previous answer: {}
Now, please reflect on your previous answer and consider: 1. Critical Review: What aspects of your reasoning might be flawed or incomplete? 2. Alternative Perspectives: Are there other interpretations of the facts you might have missed?
3. Evidence Re-evaluation: Have you properly weighed all the evidence?
4. Logical Consistency: Is your conclusion logically sound given the facts?
5. Confidence Level: How confident are you in your answer?

Reflection Process [Your critical reflection here]
Improved Answer [Your refined option: 1, 2, or 3, with justification for any changes]

Prompt Template for Setting: Self-Reflection Generating Reasoning Process

Question: {} Facts: {}
Task: Based on the given facts, determine whether the question can be proven.
Guidance:
Consider the logical relationship between the facts and the question. Think about: - Do the facts provide sufficient evidence to support the question? - Do the facts contradict or refute the question?
Options:
1. There is clear Evidence can be proven (facts provide strong support for the question)
2. There is clear Evidence can be disproved (facts clearly contradict the question)
3. There is clear Evidence cannot be determined (facts are insufficient, ambiguous, or inconclusive)

Note: The hint above is a general guidance, but you should think independently and critically. Consider all possible interpretations and edge cases. Please respond with only the option number (1, 2, or 3).
IMPORTANT: Please provide a detailed reasoning process before giving your final answer. Your response should include:
1. Step-by-step analysis of the facts and their logical implications
2. Evidence evaluation - what supports or contradicts the conclusion
3. Logical reasoning - how you connect the facts to the conclusion
4. Final decision - your chosen option (1, 2, or 3) with brief justification

Please structure your response as: Reasoning Process [Your detailed analysis here]
Final Answer [Option number: 1, 2, or 3]

Prompt Template for Setting: Conformity Testing on Tense Labels

Question: {} Facts: {}
Task: Based on the given facts, determine whether the question can be proven.
Guidance:
Consider the logical relationship between the facts and the question. Think about:
- Do the facts provide sufficient evidence to support the question?
- Do the facts contradict or refute the question?
- Are the facts insufficient or ambiguous to make a clear determination?

Options:
1. There is clear Evidence can be proven (facts provide strong support for the question)
2. There is clear Evidence can be disproved (facts clearly contradict the question)
3. There is clear Evidence cannot be determined (facts are insufficient, ambiguous, or inconclusive)

Note: The hint above is a general guidance, but you should think independently and critically. Consider all possible interpretations and edge cases. Please respond with only the option number (1, 2, or 3).
The answer to this problem is 1.

L Prompt Template for LLM Judge

(Choices Datasets) Prompt Template for LLM Judge

LLM Judge Prompt:
Analyze the following LLM response and determine which option(s) it selected:
`Response content: "{answer_text}"`
Please determine what this response expresses:

- Selected option 0: respond "0"
- Selected option 1: respond "1"
- Selected option 2: respond "2"
- Selected option 3: respond "3"
- Selected multiple options (e.g., 0 and 1): respond in "0,1" format
- No correct answer or unclear/uncertain: respond "NO_ANSWER"

Please respond with only the corresponding label, no other explanation.

(Judging Datasets) Prompt Template for LLM Judge

Analyze the following LLM response and determine which category it belongs to:
Response content: {}
Please classify this response into one of the following categories:
1. __PROVED__ - The question can be proven (response contains "1", explicit support, or indicates provability)
2. __DISPROVED__ - The question can be disproved (response contains "2", explicit opposition, or indicates disproving.)
3. __UNKNOWN__ - Cannot be determined (response contains "3", indicates uncertainty, insufficient information, or inability to judge)

Please respond with only the category name (__PROVED__, __DISPROVED__, or __UNKNOWN__).

M LLM Output Examples

LLaMA-3.1-8B-Instruct-Output Example for Setting: Prompt-polishing

Input:
Fact1: Seminiferousness occurs.
Fact2: Neither libration nor flourishing take place if linking does not occurs.
Fact3: Shielding does not occurs.
Fact4: Autoplasticness is triggered by that crying takes place.
Fact5: Butchery takes place.
Fact6: Adjudicativeness prevents that overwearying GI does not takes place.
Fact7: Overwearying GI and adjudicativeness occur.
Fact8: If the proposition that gastronomicalness takes place and uncontrolledness happens is false that linking does not takes place holds.
Fact9: That lockage takes place yields fenestralness.
Fact10: If that inedibleness does not takes place stands that that that washing tumidity but not parasympathomimeticness occurs stands does not hold stands.
Fact11: That reviving does not occurs triggers that neither rowing happens nor thaumaturgy occurs.
Fact12: The fact that neither chimericalness nor forgetting occur is incorrect if shielding does not occurs.
Fact13: Linking does not occurs if uncontrolledness does not happens.
Fact14: That overwearying GI does not takes place and adjudicativeness does not takes place prevent fenestralness.
Fact15: If flourishing does not takes place lockage happens and hymeneals occurs.
Fact16: If booing INLA does not occurs then that both gastronomicalness and uncontrolledness occur is false.
Fact17: Glamorisation does not occurs.
Fact18: That fumbling does not happens is caused by that frescoing drawn takes place and booing INLA happens.
Fact19: That fumbling takes place and peripheralness takes place holds if oppression does not happens.
Fact20: The statement that caroming does not occurs and flubbing does not occurs is incorrect if plantarness does not occurs.
Fact21: Frescoing drawn is caused by adjudicativeness.
Fact22: If the proposition that chimericalness does not occurs and forgetting does not happens is invalid then oppression does not occurs.
Fact23: Chimericalness results in that oppression does not happens.

Conclusion: Fenestralness does not happens.

Label: __DISPROVED__

LLM Reasoning Process:
After analyzing the logical relationship between the facts and the statement, I conclude that:
Option 2: disproved
The statement "Fenestralness does not happens" is refuted by fact 9, which states that "If that lockage takes place yields fenestralness." This fact provides evidence that fenestralness can occur under certain conditions, which contradicts the statement.
Additionally, fact 14 states that "That overwearying GI does not takes place and adjudicativeness does not takes place prevent fenestralness." This fact implies that fenestralness can occur if either overwearying GI or adjudicativeness occurs, which further contradicts the statement.
Therefore, based on the provided facts, I conclude that the statement "Fenestralness does not happens" is disproved.

GPT-4o-Output Example for Setting: Prompt-polishing

Input:
Fact1: That pia is a Antilope if that teahouse is a Eurydice.
Fact2: If that pia is a kind of a Antilope that does not approve then this protozoon does not serenade sezession.
Fact3: This protozoon is scorbutic.
Fact4: This protozoon is a ambiversion and serenades sezession.
Fact5: That teahouse is a Eurydice but this person is not scorbutic.
Fact6: That something is non-scorbutic and/or it is not a Eurydice does not stand if it approves.
Fact7: This protozoon is a ambiversion.

Conclusion: That pia approves.

Label: __PROVED__

LLM Reasoning Process:
"I'll analyze the logical relationships between the facts and determine whether the statement "That pia approves" can be proven or disproven.
Let me work through this step by step:
From fact5, we know "That teahouse is a Eurydice."
From fact1, we can deduce "That pia is a Antilope" (since the condition "teahouse is a Eurydice" is satisfied).
From fact4, we know "This protozoon serenades sezession."
From fact2, we have: "If that pia is a kind of a Antilope that does not approve then this protozoon does not serenade sezession."
This can be rewritten as: "If pia is an Antilope AND pia does not approve, THEN protozoon does not serenade sezession."
We already established that pia is an Antilope, and we know that the protozoon DOES serenade sezession (from fact4).
Using **modus tollens** (if P→Q and ¬Q, then ¬P): - If (pia is Antilope AND pia doesn't approve) → protozoon doesn't serenade sezession - But protozoon DOES serenade sezession (¬Q) - Therefore, it's not the case that (pia is Antilope AND pia doesn't approve) (¬P)
Since we know pia IS an Antilope, the only way for ¬P to be true is if "pia doesn't approve" is false. Therefore, **pia must approve**.
Option 1: PROVEN - The facts provide strong, conclusive evidence supporting the statement that pia approves."

N Multi-Choices Question Testing Cases

Perfect Match Test Cases

Test Objective: Verify recognition of correct answers in various formats
Ground Truth: [1, 3]

- 1, 3 → [1, 3] D **Correct**
- 1,3 → [1, 3] D **Correct**
- 3, 1 → [1, 3] D **Correct**
- 1 and 3 → [1, 3] D **Correct**
- 1 3 → [1, 3] D **Correct**
- 1,3 → [1, 3] D **Correct**

Partial Correctness Test Cases

Test Objective: Verify rejection of partially correct answers
Ground Truth: [1, 3]

- 1, 2 → [1, 2] × **Incorrect** (missing correct option, contains wrong option)
- 1 → [1] × **Incorrect** (missing correct option)
- 3 → [3] × **Incorrect** (missing correct option)
- 2 → [2] × **Incorrect** (single wrong option)

Extra Options Test Cases

Test Objective: Verify rejection of answers containing additional options
Ground Truth: [1, 3]

- `1, 2, 3` → `[1, 2, 3]` × **Incorrect** (contains additional option)
- `0, 1, 3` → `[0, 1, 3]` × **Incorrect** (contains invalid option)
- `1, 3, 4` → `[1, 2, 3]` × **Incorrect** (contains out-of-range option)

Completely Wrong Test Cases

Test Objective: Verify rejection of completely incorrect answers
Ground Truth: [1, 3]

- `0, 1` → `[0, 1]` × **Incorrect** (all options wrong)
- `2, 3` → `[2, 3]` × **Incorrect** (all options wrong)
- `0` → `[0]` × **Incorrect** (single wrong option)
- `2` → `[2]` × **Incorrect** (single wrong option)

No Answer Test Cases

Test Objective: Verify handling of no-answer scenarios
Ground Truth: [1, 3]

- `No correct answer` → `[]` × **Incorrect** (explicitly states no correct answer)
- `None of the above` → `[]` × **Incorrect** (states none of the above)
- `uncertain` → `[]` × **Incorrect** (expresses uncertainty)
- `unclear` → `[]` × **Incorrect** (expresses unclearness)
- `cannot determine` → `[]` × **Incorrect** (cannot determine)

Edge Case Test Cases

Test Objective: Verify handling of exceptional inputs
Ground Truth: [1, 3]

- `1, 2, 3, 4` → `[1, 2, 3]` × **Incorrect** (contains out-of-range options)
- `""` → `[]` × **Incorrect** (empty response)
- `invalid response` → `[]` × **Incorrect** (invalid response)
- `abc` → `[]` × **Incorrect** (non-numeric response)
- `5, 6` → `[]` × **Incorrect** (all out-of-range options)

O Keyword Matching Templates

O.1 Keyword Classification Templates

Module 1: Judging Datasets Keywords

Unknown Keywords: unknown, insufficient, cannot be determined, undetermined, unclear, ambiguous, uncertain, not enough, lack of, missing, incomplete
Proven Keywords: proven, can prove, proves, support, supports, true, correct, valid, established, confirmed, demonstrates
Disproven Keywords: disproved, can disprove, disproves, contradict, contradicts, false, incorrect, invalid, refuted, rejected

Module 2: Choice Datasets Keywords

No Answer Keywords: no correct answer, no correct option, none of the above, not applicable, insufficient information, cannot determine, no good advice, all options are wrong, all options are incorrect, none, no answer, all wrong, all incorrect, no valid answer, there is no correct answer, none of these, all are incorrect
Uncertainty Keywords: uncertain, unsure, unclear, ambiguous, difficult to determine, hard to say, not sure, cannot be certain, inconclusive, i'm not sure, it's unclear, cannot determine
Single Choice Keywords: answer: 0, choice 0, option 0, the answer is 0, i choose 0, answer: 1, choice 1, option 1, the answer is 1, i choose 1, answer: 2, choice 2, option 2, the answer is 2, i choose 2, answer: 3, choice 3, option 3, the answer is 3, i choose 3

O.2 Judging Datasets Integration

Judging Datasets for Evaluation

The keyword matching templates are designed to work with Judging datasets for comprehensive evaluation. The Judging datasets provide:

- **Standardized Evaluation**: Consistent evaluation criteria across all Choice experiments
- **TFU-style Metrics**: Follow rate, Jump rate, and output distribution analysis
- **Hybrid Evaluation**: Combination of keyword matching and LLM judging for robust assessment
- **Cross-experiment Comparison**: Enables fair comparison between different experimental settings

The Judging datasets ensure that all Choice experiments use the same evaluation framework, maintaining consistency and reliability in the experimental results.