Large Language Models Badly Generalize across Option Length, Problem Types, and Irrelevant Noun Replacements

Guangxiang Zhao*[1], Saier Hu*[2], Xiaoqi Jian*[2], Jinzhu Wu[2], Yuhan Wu[3],
Lin Sun[2†], Xiangzheng Zhang[2†]

[1]Qiyuan Tech, [2]360Zhinao, [3]Peking University

{zhaoguangxiang, yuhan.wu}@pku.edu.cn,
{husaier, jianxiaoqi, wujinzhu, sunlin1, zhangxiangzheng}@360.cn * †

Abstract

In this paper, we propose a "Generalization Stress Test" to assess Large Language Models' (LLMs) generalization ability under slight and controlled perturbations, including option length, problem types, and irrelevant noun replacements. We achieve novel and significant findings that, despite high benchmark scores, LLMs exhibit severe accuracy drops and unexpected biases (e.g., preference for longer distractors) when faced with these minor but content-preserving modifications. For example, Qwen 2.5 1.5B's MMLU score rises from 60 to 89 and drops from 89 to 36 when option lengths are changed without altering the question. Even GPT4o experiences a 25-point accuracy loss when problem types are changed, with a 6-point drop across all three modification categories. These analyses suggest that LLMs rely heavily on superficial cues rather than forming robust, abstract representations that generalize across formats, lexical variations, and irrelevant content shifts. Code can be found in: https://github.com/Qihoo360/LLMs-Generalization-Test.

1 Introduction

Large Language Models (LLMs) have achieved near-human performance across a variety of natural language processing (NLP) benchmarks, from elementary tests (Cobbe et al., 2021) to university-level challenges (Hendrycks et al., 2021). This success has spurred claims that LLMs are approaching human-like generalization capabilities (OpenAI, 2024; Bubeck et al., 2023; Jones and Bergen, 2024). However, it remains unclear whether their high benchmark scores reflect genuine generalization or if LLMs are simply exploiting superficial cues that fail under slight perturbations.

While LLMs perform well in established benchmarks, concerns have been raised about the validity of these evaluations (Chen et al., 2023; Ye et al., 2023). Data contamination, where models unintentionally learn from benchmark data included in their training, can inflate performance estimates (Brown et al., 2020; Xu et al., 2024; Ravaut et al., 2024; Zhou et al., 2023). These issues suggest that existing benchmarks have exposed patterns and may not truly assess generalization.

Recent work has focused on uncovering the actual limits of LLM generalization. One direction involves the development of dynamic evaluation methods that modify the evaluation process on the fly (Zhu et al., 2024; Yu et al., 2024), or extend the modality (Wang et al., 2025). Another approach emphasizes creating more challenging or adversarial test sets that push models beyond their current capabilities, such as MMLU-Pro (Wang et al., 2024) and GSM-Plus (Li et al., 2024a). A third line of inquiry involves introducing subtle modifications to benchmark datasets to test LLM robustness, such as altering the order of multiple-choice options or changing the format of questions (Zheng et al., 2024; Li et al., 2024b; Gupta et al., 2024; Alzahrani et al., 2024; Hong et al., 2025). While these approaches have contributed to a better understanding of LLM performance, they either totally change the original problems, increase the complexity of the evaluation, or focus on relatively limited formatting changes like option ID adjustments.

We find serious biases of recent SoTa LLMs to common patterns by introducing an evaluation framework, **Generalization Stress Tests**, which examines LLMs under three types of minor, content-preserving perturbations:

- Altering option length (e.g., increasing the length of distractors or correct options without changing their semantic content).
- Changing problem types (e.g., converting multiple-choice questions to boolean questions).

*Equal Contribution
†Corresponding Authors

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 26837–26846
November 4-9, 2025 ©2025 Association for Computational Linguistics

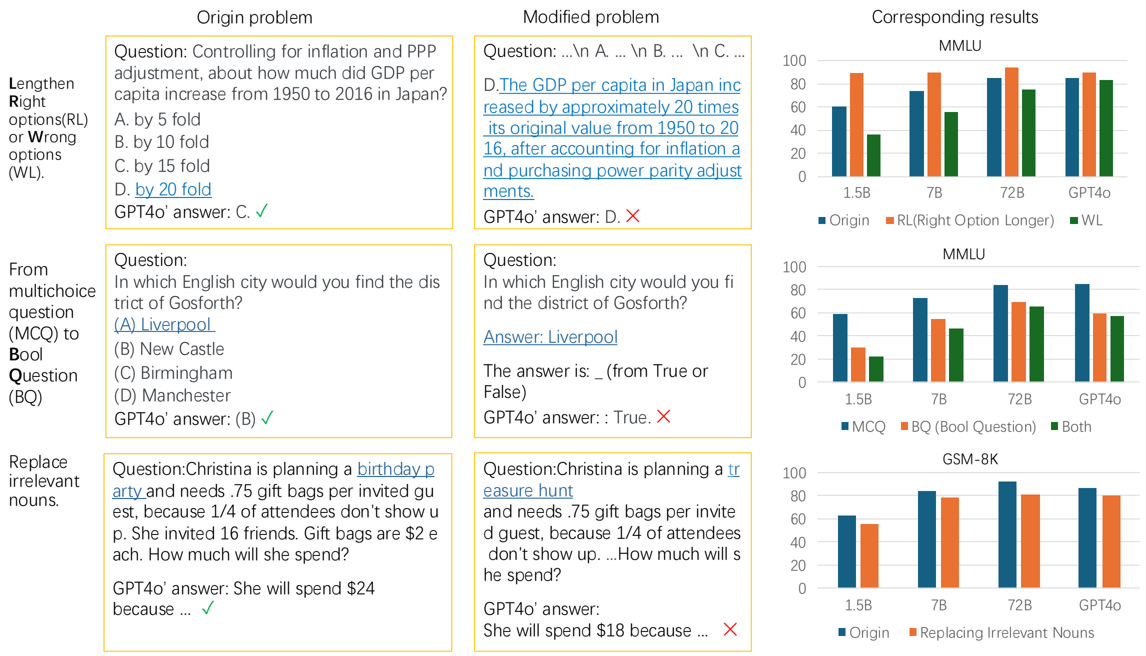

Figure 1: Generalization stress tests and summarized results. LLMs do not generalize well across various option lengths, problem types, and noun replacements. Tested models are Qwen2.5 1.5B, 7B, 72B, and GPT4o.

- Replacing irrelevant nouns (e.g., substituting semantically irrelevant nouns in prompts).

As shown in Figure 1, these simple modifications, surprisingly, lead to substantial performance degradation[1]. We observe that LLMs struggle to generalize across varying option lengths, problem types, and noun replacements. For example, Qwen 2.5 1.5B's MMLU score drops from 89 to 36 when option lengths are changed without altering the question. Even GPT4o experiences a 25-point accuracy loss when question types are changed, with a 6-point drop across all three categories. These findings reveal a critical limitation: LLMs are biased to specific irrelevant patterns and fail to replicate the human-like ability to ignore irrelevant format details.

2 Methods: Generalization Stress Tests

We conduct generalization stress tests by applying minor modifications to the original benchmark, focusing on variations in option length, scoring type, and the replacement of irrelevant nouns.

We investigate typical tasks for LLMs that include multiple-choice questions (MCQ) and open-ended question answering (Open-ended QA).

2.1 Alter Option Length to Analyze LLMs' Length Bias

Make the right option longer (RL):
Question: What is the capital of France?
A) Berlin
B) Madrid
C) Paris, a city renowned for its art, fashion, and cuisine.
D) Rome

Make one wrong option longer (WL):
Question: What is the capital of France?
A) Berlin, known for its vibrant culture and historical landmarks.
B) Madrid
C) Paris
D) Rome

Figure 2: An illustration of altering option length. The ground truth of this question is C) Paris.

To analyze whether LLMs are generalized across option length or whether LLMs are biased toward long options in MCQ. We first make all options in a problem longer by asking GPT4o[2] to make the options longer without including information that could help answer the question. Refer to Ap-

[1]We test GSM-8K for noun replacement, as some MMLU cases lack irrelevant nouns.

[2]We use its API version provided by Microsoft Azure.

pendix A for generation details.

As illustrated in Figure 2, we then design the following two types of lengthening problems: a)Make one wrong option longer (WL), b)Make the right options longer (RL).

Length Control: To assess the impact of option length on LLM generalization, we control the length of the lengthened options in the WL condition. Specifically, we ask GPT4o to generate options of varying lengths: (a) < 10 tokens, (b) 10 to 20 tokens, and (c) > 20 tokens.

Paraphrase Verification: We also enlist human experts to verify whether the paraphrased options do not introduce unintended biases or hints. Details can be found in the Appendix A.

2.2 Change Problem Type to Fairly Analyze LLMs' Scoring Bias

Cloze:
Question: What is the capital of France?
Answer: _ (*Selected from* ***whole vocabulary***)

Bool questions:
1. Question: What is the capital of France?
Answer: Paris
The answer is _(*Selected from True/False*)
2. Question: What is the capital of France?
Answer: Berlin
The answer is _(*Selected from True/False*)
Require to judge both two propositions correctly.

Figure 3: An illustration of changing the scoring type from MCQ to bool questions.

Previous work found LLMs do not generalize to different option IDs in MCQ (Zheng et al., 2024) and tried to solve this by changing the task to cloze (Alzahrani et al., 2024). However, the cloze task reduces the expected value of selecting the correct answer. Therefore, we propose changing the multiple-choice questions to Boolean questions, requiring two judgments to be accurate, so that the difficulty of the questions is as similar as possible to that of multiple-choice questions.

As illustrated in Figure 3, we derive one true proposition that concludes with the right option and one false proposition that is a randomly selected wrong option.

2.3 Replace Irrelevant Nouns to Analyze Bias towards Irrelevant Content

Problem with irrelevant noun:
Question: **John** lives in France; what is his country's capital?
A) Berlin
B) Madrid
C) Paris
D) Rome
Answer: C

Problem after modifying the irrelevant noun:
Question: **Mike** lives in France; what is his country's capital?
A) Berlin
B) Madrid
C) Paris
D) Rome
Answer: C

Figure 4: An illustration of replacing irrelevant nouns.

In open-ended QA like those in GSM8K (Cobbe et al., 2021), the questions may contain nouns that are unrelated to the answers. In this subsection, we explore the impact of changes to these unrelated nouns on the decision-making of large models. As shown in Figure 4, we replaced nouns in the questions, such as names of people and animals, ensuring that these replacements do not alter human decision-making. Details are in Appendix B.

Semantic relevance control Additionally, regarding noun replacements, we also examined the impact of the semantic proximity of the replacements. We conducted experiments in this area by instructing GPT-4o mini to perform replacements with varying degrees of semantic similarity.

3 Experiments

We perform evaluations on harness framework (Gao et al., 2024) and adopt its default setting. We evaluate models of Llama3.1 series (Dubey et al., 2024), Qwen2.5 series (Yang et al., 2024b), and GPT4o. Llama3.1, and Qwen2.5 are the most powerful small models, while GPT4o is the most powerful LLM. We evaluate LLMs on MMLU (Hendrycks et al., 2021), ARC-Challenge (Clark et al., 2018), and GSM8k (Cobbe et al., 2021). The first two are MCQ benchmarks, and the last consists of open-ended QAs. Refer to

Benchmark	Model	Origin	RL	WL
MMLU	Qwen2.5 1.5B	**60.3**	**89.0**	**36.3**
	Qwen2.5 7B	73.7	90.1	55.6
	Qwen2.5 72B	85.4	94.1	75.6
	LLaMa3.1 8B	65.5	85.6	53.6
	LLaMa3.1 70B	78.8	93.6	70.6
	GPT4o mini	76.5	87.2	70.6
	GPT4o	85.2	89.7	83.3
ARC-C	Qwen2.5 1.5B	77.3	88.9	68.1
	Qwen2.5 7B	90.0	94.3	84.0
	Qwen2.5 72B	95.8	97.2	94.4
	LLaMa3.1 8B	78.1	85.2	74.7
	LLaMa3.1 70B	91.8	96.3	90.8
	GPT4o mini	91.8	95.1	91.4
	GPT4o	96.5	97.1	95.5

Table 1: Performance on altering option length. RL refers to lengthening the right option; WL refers to lengthening the wrong option. The values are percentages.

Settings	<10	10 to 20	>20
Origin		65.5%	
RL	70.0%	75.3%	84.0%
WL	64.5%	60.7%	61.6%

Table 2: The performance of LLaMa3.1 8B on MMLU changes when gradually altering the length of correct and incorrect options.

Appendix C for detailed experimental setups.

3.1 Results of Altering Option Length

LLMs struggle to generalize across option length: From Table 1, it is evident that across all LLMs, from 1.5B to GPT4o, scores increase significantly when the length of the correct option is extended and decrease significantly when we make an incorrect option longer. Smaller models generalize even worse. In Appendix D.1, we introduce another setting of making all options longer, in which our finding that LLMs are biased towards the longer option persists.

Length matters, especially when we lengthen the right option. As shown in Table 2, changing the length can result in a difference of more than 10 points in the RL setting.

Another intriguing finding is that LLMs tend to select the right option if we make all incorrect options longer, refer to Appendix D.2.

3.2 Results of Altering Scoring Type

LLMs do not have invariant knowledge that can generalize across scoring types. As in Table 3, all models tend to score lower when the benchmarks are changed from the original format to boolean questions. Qwen2.5 1.5B and Llama3.1 8B score

Benchmark	Model	MCQ	BQ	Both
MMLU	Qwen2.5 1.5B	58.8	30.3	22.1
	Qwen2.5 7B	72.4	54.7	46.7
	Qwen2.5 72B	84.0	69.1	65.0
	LLaMa3.1 8B	64.6	40.6	32.6
	LLaMa3.1 70B	78.4	63.5	56.7
	GPT4o mini	75.1	54.5	49.2
	GPT4o	84.7	59.5	56.8
ARC-C	Qwen2.5 1.5B	74.0	40.4	35.2
	Qwen2.5 7B	89.5	69.4	66.4
	Qwen2.5 72B	95.0	85.8	84.4
	LLaMa3.1 8B	77.4	53.6	47.1
	LLaMa3.1 70B	92.1	82.7	79.2
	GPT4o mini	90.6	79.7	76.6
	GPT4o	96.2	79.6	76.2

Table 3: Performance on changing problem type from multi-choice question (MCQ) to bool questions (BQ). The values are percentages. "Both" means the percentages of examples whose MCQ and BQ are both true.

only half the points in the MMLU's "both" setting. Smaller models generalize worse.[3]

3.3 Results of Replacing Irrelevant Nouns

Models	Origin	Replace Nouns
Qwen2.5 1.5B	62.5%	54.9%
Qwen2.5 7B	83.5%	78.0%
Qwen2.5 72B	92.3%	81.9%
Llama3.1 8B	54.7%	51.7%
Llama3.1 70B	80.8%	74.2%
GPT4o mini	71.3%	64.1%
GPT4o	86.7%	79.5%

Table 4: Performance of replacing nouns on GSM8K. We report results on it since it has irrelevant nouns.

Replacing irrelevant nouns degrades performance consistently across various models. As seen in Table 5, the scores of all models drop when the terms are renamed, with the magnitude of the decrease being similar across models. GPT4o models still show a decline.

Models	Origin	High	Medium	Low
Llama3.1 8B	54.7%	51.5%	48.0%	44.0%
Qwen2.5 7B	83.5%	82.0%	78.1%	70.7%

Table 5: Model performance on replacing nouns with various semantic relevance levels.

Replacing irrelevant nouns with semantically distant words further reduces the effectiveness.

4 Discussion

[3]The "MCQ" setting is equal to "Origin" setting in Table 1, the results are slightly different since we removed the instruction "Output the answer directly" to accommodate the BQ setting.

4.1 Reasons Behind Accuracy Drops

The above ablation of results reveals that LLMs are severely biased to common but irrelevant patterns. Now, we delve a little deeper into root causes.

Could the imbalance in the test data be causing biased results? On the MMLU benchmark, a naïve policy that always selects the longest option reaches 28.3% accuracy, only +3.3% above the 25 % random baseline. This small gain shows that the length distribution of the options, by itself, is insufficient to yield the substantial performance gap reported in the paper.

Could the failures of large language models be attributed to certain mechanisms, such as the attention mechanism? Yes, we perform an analysis of attention patterns and find that lengthening options affects the attention mechanisms, causing LLMs to attend more to that option. As in Table 6, the "WL" row shows that increasing the length of option C shifts more attention toward it, confirming a length-induced bias in the attention mechanism.

Condition	A	B	C	D
Origin	0.12	0.19	**0.12**	0.21
WL	0.10	0.16	**0.26**	0.17

Table 6: The attention scores of options. The scores are from layer 0 and are averaged across all 32 heads over 24 option-orders; the numbers are the summed attention weights on the choice tokens. A is the correct answer, while C is an intentionally lengthened distractor.

4.2 Generalization of Results

Could simple interventions (e.g., fine-tuning, CoT) address these vulnerabilities? We investigate both aspects by adding simple mitigations (fine-tuning) and Chain-of-Thought (CoT) prompting. On MMLU with `Qwen2.5-7B`, fine-tuning on augmented perturbations narrows robustness variance: WL improves by $+7.8\%$ while RL decreases by -12.3% versus the base model, partially mitigating length bias. CoT further shrinks the RL–WL gap from 34.5% to 15.4%, but lowers overall accuracy (-10.7%), indicating that neither naive CoT nor simple fine-tuning fully resolves the vulnerability; stronger defenses (e.g., adversarial training or architectural changes) are likely required. CoT remains our default on GSM8K; here we additionally report MMLU results in Table 7.

Model	Origin	RL	WL
Qwen2.5-7B	73.7	90.1	55.6
Qwen2.5-7B (fine-tuned)	69.1	77.8	63.4
Qwen2.5-7B (CoT)	63.0	71.7	56.3

Table 7: Accuracy (%) on MMLU for `Qwen2.5-7B` under simple mitigations: *fine-tuning* and *Chain-of-Thought* (CoT).

Could more proprietary LLMs, such as Gemini and Claude, be fragile to these changes?

As we can see from Table 8, SoTA short CoT models, including Gemini-2.0-flash[4], Claude-3.5-sonnet[5], and Deepseek-v3 (DeepSeek-AI, 2025), are still sensitive to changes.

Model	Ori.	RL	WL	MCQ	BQ
Qwen2.5 72B	85.4	94.1	75.6	84.0	69.1
GPT4o	85.2	89.7	83.3	84.7	59.5
Deepseek-v3	86.4	90.4	84.3	85.0	71.2
Claude3.5 sonnet	86.1	92.2	85.6	87.9	30.6
Gemini2.0 flash	85.8	90.3	85.0	86.1	41.3

Table 8: Accuracy (%) of SoTA short CoT models on MMLU. 'Ori.' refers to 'Origin'.

4.3 Compare to Work in the Pre-LLM Era

Indeed, early studies have shown the limited robustness of LLMs from 2018 to 2020 (Naik et al., 2018; Zhang et al., 2019; Ribeiro et al., 2020). However, those analyses were limited to less than 1B-parameter models, whereas modern LLMs exhibit strong generalization abilities that may alter robustness patterns. Our work extends this line by examining SoTa models on comprehensive and premium-quality benchmarks and uncovers significant limitations of these capable LLMs.

4.4 Rank Changes of Models

As in D.3, the rank of models on the MMLU leaderboard changes across different settings.

5 Conclusion

This paper finds that LLMs exhibit significant performance degradation when faced with slight changes in question format, option length, or irrelevant content shifts. These findings underscore that LLMs rely on superficial patterns rather than robust, generalizable reasoning. By introducing the "Generalization Stress Tests," we offer novel understandings towards evaluating LLMs' true generalization capabilities.

[4]Gemini 2: https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/

[5]Claude 3.5 sonnet: https://www.anthropic.com/news/claude-3-5-sonnet

Limitations

This work focuses solely on non-chain-of-thought LLMs, such as GPT-4o, and does not consider emerging O1.

Ethnic Statement

This work adheres to ACL's ethical guidelines, and we state that there are no ethical concerns to our knowledge.

Acknowledgments

We thank the anonymous reviewers, the AC, SAC, and PC for their contributions to this work, and we are grateful to Change Jia for involvement in the early exploration and to Hongye Zhang for the manual annotation.

References

Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. 2024. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13787–13805, Bangkok, Thailand. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *Preprint*, arXiv:2303.12712.

Xuanting Chen, Junjie Ye, Can Zu, Nuo Xu, Rui Zheng, Minlong Peng, Jie Zhou, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. How robust is gpt-3.5 to predecessors? a comprehensive study on language understanding tasks. *Preprint*, arXiv:2303.00293.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*.

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

DeepSeek-AI. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. A framework for few-shot language model evaluation.

Vipul Gupta, David Pantoja, Candace Ross, Adina Williams, and Megan Ung. 2024. Changing answer order can decrease mmlu accuracy. *Preprint*, arXiv:2406.19470.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Mengze Hong, Wailing Ng, Chen Jason Zhang, and Di Jiang. 2025. Qualbench: Benchmarking chinese LLMs with localized professional qualifications for vertical domain evaluation. In *The 2025 Conference on Empirical Methods in Natural Language Processing*.

Cameron R. Jones and Benjamin K. Bergen. 2024. People cannot distinguish gpt-4 from a human in a turing test. *Preprint*, arXiv:2405.08007.

Qintong Li, Leyang Cui, Xueliang Zhao, Lingpeng Kong, and Wei Bi. 2024a. GSM-plus: A comprehensive benchmark for evaluating the robustness of LLMs as mathematical problem solvers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2961–2984, Bangkok, Thailand. Association for Computational Linguistics.

Wangyue Li, Liangzhi Li, Tong Xiang, Xiao Liu, Wei Deng, and Noa Garcia. 2024b. Can multiple-choice questions really be useful in detecting the abilities of LLMs? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2819–2834, Torino, Italia. ELRA and ICCL.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Mathieu Ravaut, Bosheng Ding, Fangkai Jiao, Hailin Chen, Xingxuan Li, Ruochen Zhao, Chengwei Qin, Caiming Xiong, and Shafiq Joty. 2024. How much are large language models contaminated? a comprehensive survey and the llmsanitize library. *Preprint*, arXiv:2404.00699.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Xiaonan Wang, Bo Shao, and Hansaem Kim. 2025. Toward reliable vlm: A fine-grained benchmark and framework for exposure, bias, and inference in korean street views. *Preprint*, arXiv:2506.03371.

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. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Preprint*, arXiv:2406.01574.

Cheng Xu, Shuhao Guan, Derek Greene, and M-Tahar Kechadi. 2024. Benchmark data contamination of large language models: A survey. *Preprint*, arXiv:2406.04244.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024a. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024b. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *Preprint*, arXiv:2303.10420.

Zhuohao Yu, Chang Gao, Wenjin Yao, Yidong Wang, Wei Ye, Jindong Wang, Xing Xie, Yue Zhang, and Shikun Zhang. 2024. KIEval: A knowledge-grounded interactive evaluation framework for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5967–5985, Bangkok, Thailand. Association for Computational Linguistics.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023. Don't make your llm an evaluation benchmark cheater. *Preprint*, arXiv:2311.01964.

Kaijie Zhu, Jiaao Chen, Jindong Wang, Neil Zhenqiang Gong, Diyi Yang, and Xing Xie. 2024. Dyval: Dynamic evaluation of large language models for reasoning tasks. In *The Twelfth International Conference on Learning Representations*.

A Prompts and Verification in Altering Option Length

A.1 Prompts

We chose the GPT-4o to lengthen options.

The default prompt to lengthen options is: The user will give you a question, the choices, and the answer from a dataset. Rewrite the four choices into longer ones. Make sure not to change the question willingly. Make sure that the rewritten options do not contain a hint of the correct answer.

The prompt to control option length is: We concatenate the default prompt to one of the following prompts.

- Make sure that each rewritten option contains no more than 10 words.
- Make sure that each rewritten option at least 10 words and no more than 20 words.
- Make sure that each rewritten option contains at least 20 words.

We set the temperature to 0, and the other setting is the same as the default.

A.2 Verification Process

We manually verified the rewritten sentences to check whether lengthening the sentence introduced factors related to the answer or changed the question's meaning. We manually checked 100 examples from MMLU and found that 99 had no issues, while 1 changed the original meaning of the question. The rewriting accuracy was 99%.

B Prompts in Replacing Irrelevant Nouns

We found that GPT-4o and GPT-4o mini perform similarly on this task. To reduce carbon emissions, we chose the GPT-4o mini.

The prompt to simply replace irrelevant nouns is: Assist in creatively substituting nouns in mathematical problems to prevent students from memorizing solutions. The replacements should be imaginative, ensuring the mathematical relationships and the accuracy of the solutions are preserved. “input_text” Other than replacing nouns, do not alter the original word order sentence structure, or add or remove any sentences. Give the modified question directly.

The prompt to alter semantic relevance is: Substitute nouns and some relevant words in the mathematical problems creatively to prevent students from memorizing solutions. The replacements should be done in three levels:

- Level 1: Only replace nouns with semantically similar words (e.g., 'apple' becomes 'banana').
- Level 2: Replace nouns and verbs with words that differ in meaning but are still within the realm of common sense (e.g., 'apple' becomes 'elephant', 'eat fruit' becomes 'drink coke').

Bench	Model	Ori.	AL	RL	WL
MMLU	Qwen2.5 1.5B	**60.3**	**54.7**	**89.0**	**36.3**
	Qwen2.5 7B	73.7	69.2	90.1	55.6
	Qwen2.5 72B	85.4	81.3	94.1	75.6
	LLaMa3.1 8B	65.5	64.3	85.6	53.6
	LLaMa3.1 70B	78.8	76.0	93.6	70.6
ARC-C	Qwen2.5 1.5B	77.3	67.3	88.9	68.1
	Qwen2.5 7B	90.0	85.3	94.3	84.0
	Qwen2.5 72B	95.8	93.1	97.2	94.4
	LLaMa3.1 8B	78.1	78.6	85.2	74.7
	LLaMa3.1 70B	91.8	89.9	96.3	90.8

Table 9: Performance on altering option length. AL refers to lengthening all options. RL refers to lengthening the right option. WL refers to lengthening the wrong option. The values are percentages.

- Level 3: Replace words as much as possible with highly imaginative and fantastical words, if you think it still makes sense in mathematical problems. (e.g., 'apple' becomes 'alien gemstone').

Apart from replacing nouns and some relevant words, maintain the original word order, sentence structure, and do not add or remove any sentences. Give three modified sentences directly, one for each level, only separated by '###'. Don't return anything else including 'Level 1', 'Level 2', 'Level 3' but only "###". This is the original question: input_text

We set temperature to 0.1, top-p to 1, top-k to 0, and repetition_penalty to 0.

C Experiment Setup Details

This section outlines the foundational setup of our experiments and analyses, including the evaluation framework and methods employed, as well as the benchmarks and models evaluated.

C.1 Evaluation Protocol

We perform evaluations on the Harness framework (Gao et al., 2024). We chose Harness because it is a flexible, configurable, reproducible framework. Unless specified, we follow the default parameter of the harness. Unless otherwise specified, our evaluations are conducted in a 5-shot manner, with few-shot examples drawn from the benchmarks' corresponding training sets.

C.2 Models

We evaluate models of Llama3.1 series (Dubey et al., 2024), Qwen2 series (Yang et al., 2024a), and GPT4o. Llama3.1 and Qwen2.5 are the most powerful small models, while GPT4o is the most powerful LLM. We list all models below.

- Llama3.1 8B, Llama3.1 70B;
- Qwen2.5 1.5B, Qwen2.5 7B, Qwen2.5 72B;
- GPT4o, GPT4o mini.

C.3 Benchmarks

We evaluate LLMs on MMLU, ARC, Helaswag, GSM-MCQ, and GSM8k. The first four are MCQ benchmarks, and the last consists of open-ended questions.

- **MMLU** (Hendrycks et al., 2021) is a multi-task benchmark that covers 57 tasks ranging from elementary to college level. These tasks cover multiple disciplines, e.g., math, physics, law, history, etc. The whole test set consists of 14,042 examples. Following common practice, we calculate the accuracy of each task and report the average score across all tasks.
- **ARC** (Clark et al., 2018) is also a multitask dataset that includes data from eight types of tasks, testing aspects such as common sense, multi-hop reasoning, and algebraic operations, with 3,548 samples. ARC has two subsets: one is ARC-Challenge (abbreviated as ARC-C), and the other is ARC-Easy (abbreviated as ARC-E). The challenge set includes only those data that cannot be answered through retrieval and word co-occurrence methods, making it more difficult.
- **GSM-8K** (Cobbe et al., 2021) examines multi-step math word problems, which are relatively easy and designed to be solvable by middle school students. GSM8K is presented in an open-ended question format, unlike multiple-choice questions. It consists of 1,319 test questions.

C.4 Budget

We performed experiments with an H800 GPU; the total cost of the experiments was about 1000 GPU hours.

C.5 Random Seeds

All reported numbers are from a single wrong, since the model is deterministic with our default decoding temperature $T = 0$, thus, changing seeds has no effect. We nevertheless tested robustness at $T = 0.5$ and $T = 1.0$ while varying Python/NumPy/Torch seeds; results for Qwen-2.5 7B on MMLU are shown below. The RL/WL gap persists, and our conclusions remain unchanged.

Temp.	Seed	Origin	RL	WL
0	0, 1234, 1234	73.7	90.1	55.7
	1, 11, 111	73.7	90.1	55.7
	2, 22, 222	73.7	90.1	55.7
0.5	0, 1234, 1234	71.6	88.2	54.9
	1, 11, 111	71.5	88.5	54.9
	2, 22, 222	71.5	88.3	55.3
1.0	0, 1234, 1234	67.9	83.4	52.2
	1, 11, 111	67.3	83.8	51.9
	2, 22, 222	67.2	83.8	53.0

Table 10: Model accuracy (%) for Qwen-2.5 7B on MMLU under different temperatures and seeds. Numbers are averaged over three runs when applicable; at $T = 0$ the model is deterministic. Each seed setting contains three seeds regarding Python, NumPy, and Torch.

D Additional Results

D.1 Making All Options longer

We can see from Table 9 that LLaMa is more robust than Qwen, and larger models are more robust than smaller models, when we make all options longer. Besides, even if we introduce the setting of AL, our conclusion that LLMs are vulnerable to option lengths and biased to long options is not changed.

D.2 Make All Wrong Options Longer

Model	origin	WL	WL-ALL
Llama3.1 8B	65.5%	53.6%	64.8%
Llama3.1 70B	78.8%	70.6%	82.4%
gpt-4o	85.2%	83.3%	85.6%

Table 11: Results of making all wrong options longer on the MMLU benchmark.

Making all wrong options could expose the right answer. From Table 11, we can see that if all the incorrect options are lengthened, the model will choose the only correct option that hasn't been lengthened.

D.3 Rank Changes of Models

Note of Kendall τ . Kendall τ is a non-parametric rank-correlation coefficient that measures how similarly two lists are ordered by comparing the balance of concordant versus discordant item pairs, ranging from -1 (complete disagreement) to $+1$ (perfect agreement).

Table 12: Accuracy (%), rank, and rank changes of models on MMLU under different protocols.

Model	Origin		RL			WL		
	Score	Rank	Score	Rank	ΔRank	Score	Rank	ΔRank
Qwen2.5 1.5B	60.3	7	89.0	5	↑ 2	36.3	7	—
Qwen2.5 7B	73.7	5	90.1	3	↑ 2	55.6	5	—
Qwen2.5 72B	85.4	1	94.1	1	—	75.6	2	↓ 1
LLaMa 3.1 8B	65.5	6	85.6	7	↓ 1	53.6	6	—
LLaMa 3.1 70B	78.8	3	93.6	2	↑ 1	70.6	3	—
GPT-4o mini	76.5	4	87.2	6	↓ 2	70.6	3	↑ 1
GPT-4o	85.2	2	89.7	4	↓ 2	83.3	1	↑ 1
Kendall τ	—		0.52			0.88		

Table 13: Accuracy (%), rank, and rank changes of models on MMLU (MCQ vs. BQ).

Model	MCQ		BQ		
	Score	Rank	Score	Rank	ΔRank
Qwen2.5 1.5B	58.8	7	30.3	7	—
Qwen2.5 7B	72.4	5	54.7	4	↑ 1
Qwen2.5 72B	84.0	2	69.1	1	↑ 1
LLaMa 3.1 8B	64.6	6	40.6	6	—
LLaMa 3.1 70B	78.4	3	63.5	2	↑ 1
GPT-4o mini	75.1	4	54.5	5	↓ 1
GPT-4o	84.7	1	59.5	3	↓ 2
Kendall τ	—		0.71		

As we can see, the rank of the LLMs changes when we apply the perturbation. The Kendall τ between the WL and RL rankings is 0.39; since a Kendall τ below 0.4 is not considered strongly correlated, **switching the evaluation protocol still has a pronounced influence on the rank shifts**.