WiCkeD: A Simple Method to Make Multiple Choice Benchmarks More Challenging

Ahmed Elhady[1] **Eneko Agirre**[1] **Mikel Artetxe**[1,2]
[1]HiTZ Center, University of the Basque Country (UPV/EHU) [2]Reka AI
{ahmed.salemmohamed,e.agirre,mikel.artetxe}@ehu.eus

Abstract

We introduce WiCkeD, a simple method to increase the complexity of existing multiple-choice benchmarks by randomly replacing a choice with *"None of the above"*, a method often used in educational tests. We show that WiCkeD can be automatically applied to any existing benchmark, making it more challenging. We apply WiCkeD to 7 popular benchmarks and use it to evaluate 18 open-weight LLMs. The performance of the models drops 12.1 points on average with respect to the original versions of the datasets. When using chain-of-thought on 3 MMLU datasets, the performance drop for the WiCkeD variant is similar to the one observed when using the LLMs directly, showing that WiCkeD is also challenging for models with enhanced reasoning abilities. WiCkeD also uncovers that some models are more sensitive to the extra reasoning required, providing additional information with respect to the original benchmarks. We release our code and data at https://github.com/ahmedselhady/wicked-benchmarks.

1 Introduction

Multiple choice question (MCQ) benchmarks are widely used to evaluate Large Language Models (LLMs). This format consists of a question and a limited set of options, which include a correct (or best) answer and several distractors that are either incorrect or less appropriate (see Figure 1). There are various MCQ datasets that focus on different capabilities, including factual knowledge and reasoning as in MMLU (Hendrycks et al., 2021) and Arc-challenge (Clark et al., 2018), common sense as in Commonsense-QA (Talmor et al., 2019), truthfulness as in TruthfulQA (Lin et al., 2022), and domain-specific knowledge (Alonso et al., 2024; Hosseini et al., 2024). Unfortunately, most of these benchmarks got quickly saturated in the recent era dominated by LLMs, motivating harder datasets to better gauge the abilities of newer models. However, developing benchmarks is a laborious and expensive process.

Motivated by this, several recent works have explored strategies to make existing benchmarks harder, which can serve as an alternative to creating new benchmarks from scratch. For example, Gema et al. (2024) identified erroneous questions in the MMLU benchmark, and re-annotated 3k questions to be harder and more robust. Similarly, Wang et al. (2024) presented MMLU-Pro, a harder version of the MMLU benchmark that replaces noisy questions with harder ones and expands the number of distractors to include more plausible yet incorrect ones. While increasing the number of distractors reduces the probability of correct guesses by chance, creating plausible and coherent distractors is challenging and often requires manual verification (McIntosh et al., 2024).

In this work, we propose a simple yet effective method to make existing benchmarks more challenging without the need to add distractors. Namely, we present the **Wi**ld-**C**ard **D**istractor (WiCkeD) which creates a variant of any existing MCQ benchmark by keeping the question unchanged, and randomly replacing one of the choices with a wild-card distractor, *None of the above* (see Figure 1). We create WiCkeD variants of 7 popular benchmarks, and use them to evaluate 18 open-weight LLMs varying in size, model family, and training recipe. The WiCkeD datasets suffer a performance drop of 7.2-19.7 points with respect to the original datasets, depending on the model being evaluated. Using chain-of-thought does not prevent the drop (1.4-14.6), showing that WiCkeD can be used to assess reasoning capabilities. The large variance across models shows that WiCkeD is not only challenging, but it also uncovers differences in model capabilities that are not captured by the original benchmarks.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 1183–1192

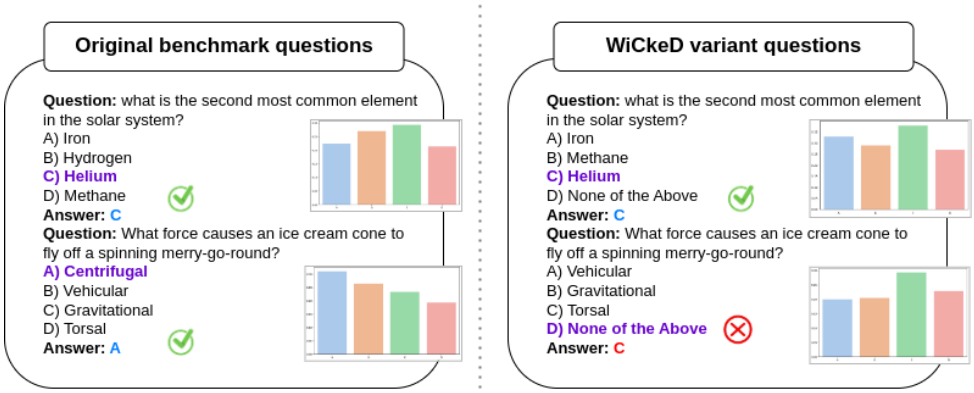

Figure 1: Two samples from MMLU-Pro (left) and its WiCkeD variant (right), where *Hydrogen* and *Centrifugal* were removed. Correct answers in bold. Llama-3.1 8B correctly answers both original questions but fails on the WiCkeD variant for the second question. The probability distribution of the model for each answer is also shown.

2 Related Work

2.1 Challenges in LLMs MCQ Benchmarks

Several works raised concerns about the effectiveness of MCQ benchmarks in LLM assessment. For example, Balepur et al. (2024) showed that some LLMs can answer MCQs using only the answer choices, without seeing the questions, and perform well-above baselines. Furthermore, more works suggested that LLMs are biased towards certain answer keys (A/B/C/D) due to unbalanced prior probabilities rather than actual knowledge (Myrzakhan et al., 2024; Clark et al., 2018).

Another line of research attributes LLMs hallucinations to being unable to identify when they lack sufficient knowledge about the subject matter (Li et al., 2024; Ji et al., 2022). Nonetheless, current evaluation benchmarks do not assess this capability effectively. A third line argues that the overestimation in performances of models is due to contamination (i.e. the model is trained on the test sets of the benchmarks), leading to wrong conclusions (Golchin and Surdeanu, 2024; Dong et al., 2024; Sainz et al., 2023). We view our work as an addition towards efficient evaluation of LLMs to avoid spurious correlations and account for knowledge and reasoning gaps.

2.2 None of the Above in Educational Tests

Multiple-choice questions (MCQs) are effective assessments when they include plausible distractors, as they encourage deeper processing to think not only about why a given choice is correct, but also why other choices are wrong and improve knowledge recall (Little et al., 2019; Little and Bjork, 2015). The use of *None of the above* as a distractor in MCQs is an area of research and debate. It can provide unique insight into the understanding of the examinees and potentially differentiate their abilities (David DiBattista and Fortuna, 2014; Dochy et al., 2001). However, *None of the above* can affect the confidence of the examinee, leading them to avoid selecting *None of the above* as the correct answer, even when it is true (Little, 2023; Odegard and Koen, 2007). Nevertheless, incorporating *None of the above* into practice tests can enhance the learning process by encouraging deeper engagement with the material (David DiBattista and Fortuna, 2014; Pezeshkpour and Hruschka, 2024; Zheng et al., 2024).

3 Methodology

We propose a method to automatically create a more challenging version of any existing MCQ benchmark without requiring any manual annotation. The difficulty of MCQ has been linked to the reasoning necessary to discriminate between competing options (McIntosh et al., 2024; Wang et al., 2024). We hypothesize that detecting the absence of the correct answer within the provided options is more challenging than selecting the correct one. To that end, we propose to add a wild-card choice *None of the above*. Note that adding *None of the above* as an additional option would not make sense, as the correct answer is always the correct option, we thus propose to replace one of the options instead.

3.1 The WiCkeD Algorithm

Given a benchmark that consists of M examples where each has N choices (one correct answer and $N-1$ distractors), we uniformly sample one option

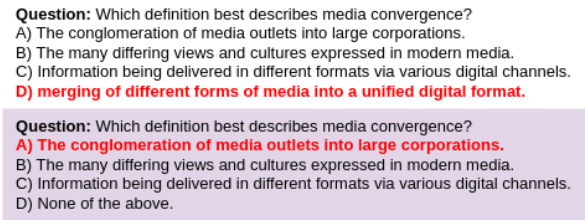

Figure 2: Applying WiCkeD on a single best answer (SBA) example (best answer D, second best answer A) would lead to an incoherent WiCkeD variant (incorrectly having *None of the above* as the gold correct answer instead of A). We thus copy SBA examples verbatim, see § 3.2 for details.

to be omitted, and append the wildcard option *None of the above*. When the correct option is replaced, the new correct option is *None of the above*. When a distractor option is replaced, the correct option continues to be correct. Figure 1 shows the result of applying WiCkeD to two examples. The goal is to produce a variant for each benchmark that contains the same number of M examples.

3.2 Coherence of WiCkeD Examples

The above algorithm does not always produce coherent examples. In some cases, there are more than one correct candidate, but only one of them is the most appropriate (see Figure 2, where D is the best answer and A is the second best answer). With the above procedure, when the replaced option is the correct one (e.g. option D in the figure), the WiCkeD variant would add *None of the above* and take this option as the correct one. However, this would be incoherent, because having removed D, A becomes the next best option. We call these examples Single Best Answer (SBA) as opposed to Single Correct Answer (SCA, where the distractors are all incorrect). As we want to keep the same number of examples we avoid adding *None of the above* to SBA examples and copy them unchanged to the WiCkeD variant of the benchmark.[1]

In order to train an example classifier to detect SBA examples, we selected four representative benchmarks (MMLU, MMLU-Pro, Truthful-QA and Commonsense-QA), sampled 4000 examples, and split them into evaluation (25%) and train (75%). We used GPT-4o-mini to automatically label the examples as SBA or SCA, and further annotated the evaluation split manually. Given the cost and slow speed of GPT-4o-mini, we used the synthetic labels to train a classifier based on BERT[2] (Devlin et al., 2019).

The recall on SBA examples for the classifier is over 98.9%, showing that we are able to detect nearly all SBA examples, and would thus have 1.1% noisy WiCkeD examples (that is, examples in the benchmark that have *None of the above* as the correct option even if a correct option exists). See Appendix A for more details about the training and evaluation procedure.

4 Experimental Setup

4.1 Benchmarks

We apply WiCkeD to six popular MCQ benchmarks that assess the knowledge, language comprehension, reasoning, and truthfulness of LLMs: MMLU, MMLU-Pro, MMLU-Redux, GPQA, CommonsenseQA, Truthful-QA, and Arc-challenge. To ensure reproducibility, we use Eval-Harness (Gao et al., 2024). Given that the selection of the option to be replaced is random, we generate five WiCkeD variants for each benchmark, and report mean and standard deviation.

Regarding the amount of SBA examples, MMLU, MMLU-Redux and MMLU-pro have the largest amount (~ 20%), with the rest of the benchmarks having less than 5% (see Appendix A). SBA examples are copied verbatim to the WiCkeD variants, but the fact that at least 80% of the examples are effectively altered makes the WiCkeD variants significantly more challenging, as we will see. Other benchmarks have less than 5% SBAs; we also leave them unchanged.

4.2 Models

We evaluate WiCkeD on 18 open-weight models covering different families and sizes. Namely, we evaluate the base and instruction-tuned models of Qwen2.5 7B, 14B and 72B (Qwen et al., 2025), Llama3.1 8B and 70B (Grattafiori et al., 2024), Gemma2 9B and 27B (Riviere et al., 2024), and Mistral-7B (Jiang et al., 2023). We also selected two DeepSeek-R1 models for their improved reasoning capabilities: distill-Lllama3.1-8B and distill-Qwen7 (DeepSeek-AI et al., 2025).

The LLM models are evaluated on the benchmarks following the standard multiple-choice prompting procedure (Robinson et al., 2023), see

[1] Excluding the SBA questions amplifies the impact of WiCkeD yet adds no other signals about performance, therefore, we leave them unchanged.

[2] https://huggingface.co/ahmedselhady/bert-base-uncased-sba-clf

Model	Size	IT	Original	WiCkeD	Δ
DS-R1-Llama	8B	-	56.6	48.6	-7.9 ±1.1%
DS-R1-Qwen	7B	-	60.8	53.4	-7.3 ±1.6%
Llama-3.1	8B	-	61.4	52.2	-9.2 ±1.7%
	8B	✓	66.0	55.0	-11.0 ±0.9%
	70B	-	76.8	67.0	-9.8 ±2.1%
	70B	✓	77.1	64.5	-12.6 ±1.3%
Mistral	7B	-	59.8	46.5	-13.2 ±1.2%
	7B	✓	59.0	47.2	-11.8 ±1.1%
Qwen-2.5	7B	-	74.7	54.9	-19.7 ±1.5%
	7B	✓	73.5	59.0	-14.5 ±1.3%
	14B	-	78.9	66.3	-12.6 ±2.1%
	14B	✓	78.9	66.6	-12.3 ±1.8%
	72B	-	84.6	72.6	-12.0 ±0.9%
	72B	✓	82.6	69.3	-13.3 ±1.0%
Gemma-2	9B	-	67.3	56.3	-10.9 ±1.2%
	9B	✓	73.3	57.6	-15.7 ±1.2%
	27B	-	68.0	54.6	-13.4 ±2.0%
	27B	✓	74.8	61.9	-12.9 ±2.3%
Average			**70.8**	**58.5**	**-12.2** ±1.5%

Table 1: Average performance on original and WiCkeD variants of the six benchmarks. IT: instruction-tuned. **Δ**: degradation from original performance

Size Group	Original	WiCkeD	Δ
≤10B	63.4	51.9	-11.5
	67.9	54.7	-13.2
10-70B	73.5	60.5	-13.0
	76.8	64.2	-12.6
≥70B	80.7	69.8	-10.9
	79.8	66.9	-12.9

Table 2: Average performance per size group in original and WiCkeD variants of the six benchmarks using direct prompting. The degradation (Δ) does not decrease with model size scaling.

Appendix C. We set the number of few-shot examples to five, in order to ensure that in most cases there is at least one example where *None of the above* is the correct option.

In addition, we also evaluate the LLM models using zero-shot chain-of-thoughts prompting (CoT) on the three benchmarks commonly used to assess the reasoning capabilities of LLMs: MMLU, MMLU-Pro, MMLU-Redux, and GPQA. We also include 3 state-of-the-art closed-source models: OpenAI GPT-4o and GPT-4o-mini (OpenAI et al., 2024) and Gemini Flash 2.0.[3] We set the maximum generation length to 4096, unless limited by the model itself.

5 Results and Discussion

5.1 Main Results

Table 1 shows the mean accuracy of the models on the original and WiCkeD benchmarks, with a significant drop in performance. Qwen2.5-7B suffers the largest degradation (19.73%), while its DeepSeek-R1 distilled version (DeepSeek-R1-Qwen7B) suffers the least (7.35%). This suggests that models with better reasoning capabilities, like R1, are better equipped to deal with the added complexity.

Prominently, the WiCkeD variants shuffle the ranking of models. For example, the Qwen2.5-7B and Qwen2.5-7B-IT models originally performed close to the Llama-3.1-70B model. However, on the WiCkeD variants, they lag behind it by 12.1% and 8%, respectively. Similar patterns can be seen in Gemma-2-9B-IT and Gemma-2-27B-IT, which lag behind Llama-3.1-70B by 9.5% and 5.3%, respectively. Qwen2.5-72B and Llama-3.1-70B are the models that perform best in WiCkeD. There is no clear advantage from instruction-tuning, as results vary depending on the model family. Furthermore, the degradation does not decrease with model size scaling, as shown in Table 2.

5.2 Chain-of-Thought Results

Table 3 shows the performance of the models[4] on the MMLU, MMLU-pro, MMLU-Redux, and GPQA WiCkeD benchmarks. The drop for these three benchmarks without CoT (direct columns in the table) is lower than the other three benchmarks, but applying CoT does not reduce the drop in WiCkeD variants, which stays above 5%. This is remarkable given that CoT is very effective at improving results on MMLU and related benchmarks. Instruction-tuned models experience significantly less degradation than their base models, especially when using CoT (see Appendix B for additional details). Notably, the DeepSpeed-R1 distilled models, Qwen7B and Llama3.1-8B, suffer around 2% each. Similarly, instruction-tuned Qwen2.5 7B and 14B suffer less than 2%. We hypothesize this is due to their enhanced reasoning capabilities.

As shown in Table 4, popular closed-source models also suffer a relatively large drop in the WiCkeD variants (3.9-8.2%). This suggests that state-of-the-art commercial models are also susceptible to the phenomenon studied in our work.

[3] deepmind.google/technologies/gemini/flash-lite/

[4] Due to compute constraints, we could not run CoT for the ~70B models

			Direct		CoT	
Model	**Size**	**IT**	**WiCkeD**	**Δ**	**WiCkeD**	**Δ**
DS-R1-Llama	8B	-	30.3	-4.1	80.1	-2.0
DS-R1-Qwen	7B	-	30.6	-4.3	74.9	-2.5
Llama-3.1	8B	-	39.7	-3.2	53.9	-5.8
	8B	✓	43.6	-2.7	57.2	-3.4
Mistral	7B	-	35.9	-3.4	36.3	-11.6
	7B	✓	33.5	-5.7	43.8	-4.9
Qwen-2.5	7B	-	45.5	-6.9	43.0	-14.6
	7B	✓	47.1	-5.3	55.4	-1.7
	14B	-	55.6	-3.6	61.5	-3.97
	14B	✓	56.7	-3.4	64.0	-1.4
Gemma-2	9B	-	36.1	-12.2	41.2	-8.9
	9B	✓	44.1	-9.3	56.3	-4.4
	27B	-	36.1	-10.8	59.2	-4.1
	27B	✓	51.3	-3.8	60.3	-3.8
Avg			**41.9**	**-5.6**	**56.3**	**-5.2**

Table 3: Performance on WiCkeD variants for MMLU, MMLU-pro, and MMLU-Redux with and without CoT. IT: instruction-tuned. **Δ**: degradation from the original benchmark.

6 Conclusion

In this paper, we introduced a simple automatic method to create more challenging variants from an existing MCQ benchmark. The large drop in the results shows that WiCkeD challenges the knowledge and reasoning of LLMs, as they need to identify the absence of the correct answer, even when using CoT. We showed that models with better reasoning capabilities suffer less in WiCkeD, such as the original Qwen7B and its distilled version of DS-R1. We see WiCkeD as an addition towards efficient evaluation of LLMs to avoid spurious correlations and challenge reasoning and knowledge gaps. A deeper look into why some models are more sensitive to WiCkeD than others can provide significant insights about uncovered limitations. We release all the code and data under open licenses.

Limitations

We manually confirmed the applicability of WiCkeD on some popular multiple-choice benchmarks whose questions can be categorized into SBAs and SCAs. However, for other benchmarks, WiCkeD might need further verification.

Acknowledgments

This work has been partially supported by the Basque Government (Research group funding IT1805-22). Ahmed Elhady holds a PhD grant supported by the Basque Government (IKER-GAITU project).

Model	**WiCkeD**	**Δ**
GPT-4o	81.4	-3.9
GPT-4o-mini	74.6	-8.2
Gemini Flash 2.0	80.3	-7.8
Avg	**78.8**	**-6.6**

Table 4: Performance of closed-source models on WiCkeD variants for MMLU, MMLU-pro, and MMLU-Redux using CoT. **Δ**: degradation from the original benchmark.

References

Iñigo Alonso, Maite Oronoz, and Rodrigo Agerri. 2024. Medexpqa: Multilingual benchmarking of large language models for medical question answering. *Artificial Intelligence in Medicine*, 155:102938.

Nishant Balepur, Abhilasha Ravichander, and Rachel Rudinger. 2024. Artifacts or abduction: How do llms answer multiple-choice questions without the question? *Preprint*, arXiv:2402.12483.

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. *Preprint*, arXiv:1803.05457.

Jo-Anne Sinnige-Egger David DiBattista and Glenda Fortuna. 2014. The "none of the above" option in multiple-choice testing: An experimental study. *The Journal of Experimental Education*, 82(2):168–183.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Filip Dochy, George Moerkerke, Erik De Corte, and Mien Segers. 2001. The assessment of quantitative problem-solving skills with "none of the above"-items (nota items). *European Journal of Psychology of Education*, 16:163–177.

Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. 2024. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12039–12050, Bangkok, Thailand. Association for Computational Linguistics.

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, and 5 others. 2024. A framework for few-shot language model evaluation.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile van Krieken, and Pasquale Minervini. 2024. Are we done with mmlu? *Preprint*, arXiv:2406.04127.

Shahriar Golchin and Mihai Surdeanu. 2024. Time travel in LLMs: Tracing data contamination in large language models. In *The Twelfth International Conference on Learning Representations*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Pedram Hosseini, Jessica M. Sin, Bing Ren, Bryceton G. Thomas, Elnaz Nouri, Ali Farahanchi, and Saeed Hassanpour. 2024. A benchmark for long-form medical question answering. *Preprint*, arXiv:2411.09834.

Yunjie Ji, Liangyu Chen, Chenxiao Dou, Baochang Ma, and Xiangang Li. 2022. To answer or not to answer? improving machine reading comprehension model with span-based contrastive learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1292–1300, Seattle, United States. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Moxin Li, Wenjie Wang, Fuli Feng, Fengbin Zhu, Qifan Wang, and Tat-Seng Chua. 2024. Think twice before trusting: Self-detection for large language models through comprehensive answer reflection. *Preprint*, arXiv:2403.09972.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. *Preprint*, arXiv:2109.07958.

Jeri L Little. 2023. Does using none-of-the-above (nota) hurt students' confidence? *Journal of Intelligence*, 11(8):157.

Jeri L Little and Elizabeth Ligon Bjork. 2015. Optimizing multiple-choice tests as tools for learning. *Memory & cognition*, 43:14–26.

Jeri L Little, Elise A Frickey, and Alexandra K Fung. 2019. The role of retrieval in answering multiple-choice questions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 45(8):1473.

Timothy R. McIntosh, Teo Susnjak, Nalin Arachchilage, Tong Liu, Paul Watters, and Malka N. Halgamuge. 2024. Inadequacies of large language model benchmarks in the era of generative artificial intelligence. *Preprint*, arXiv:2402.09880.

Aidar Myrzakhan, Sondos Mahmoud Bsharat, and Zhiqiang Shen. 2024. Open-llm-leaderboard: From multi-choice to open-style questions for llms evaluation, benchmark, and arena. *Preprint*, arXiv:2406.07545.

Timothy N Odegard and Joshua D Koen. 2007. "none of the above" as a correct and incorrect alternative on a multiple-choice test: Implications for the testing effect. *Memory*, 15(8):873–885.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, and 178 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Joshua Robinson, Christopher Michael Rytting, and David Wingate. 2023. Leveraging large language

models for multiple choice question answering. *Preprint*, arXiv:2210.12353.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. *Preprint*, arXiv:2309.03882.

A Detecting Single Best Answer examples

To ensure the reliability of the automatic identification of single-best-answer (SBA) questions, we uniformly sample 4K questions from the MMLU, MMLU-Pro, Commonsense-QA, and Truthful-QA benchmarks, which we divide into 1K and 3K splits. We then manually annotate the 1K samples and optimize GPT-4o-mini prompt on them for best recall. Table 6 shows the prompt template for GPT-4o-mini, which we used to annotate the 4K questions. The 3K split was then used to train our Bert-based SBA classifier on them. The classifier was trained for 2 epochs, using a learning rate of 1e-04. The model was frozen, except for the last layer and the classification head.

Table 5 shows the percentages of SBA questions on the 1K split as determined by our manual annotations, GPT-4o-mini, and the SBA classifier. The classifier is the preferred one, as it is the most conservative, that is, it detects the most SBA examples, which would be copied verbatim to the WiCkeD variant of the benchmark. The evaluation figures in the table confirm this choice, as the classifier has higher recall. The small drop in precision is harmless, as it means that we will not add *None of the above* option to those examples, and will be copied verbatim. In other words, we can estimate that WiCkeD contains 1% of incoherent examples (where there is a valid option even if *None of the above* is recorded as the correct option), and 5% of examples which do not have a *None of the above* option even if we could have added it if the classifier had 100% precision. These figures confirm the high quality of the WiCkeD variants. Table 7 shows the final SBA percentages for each benchmark as determined by the classifier.

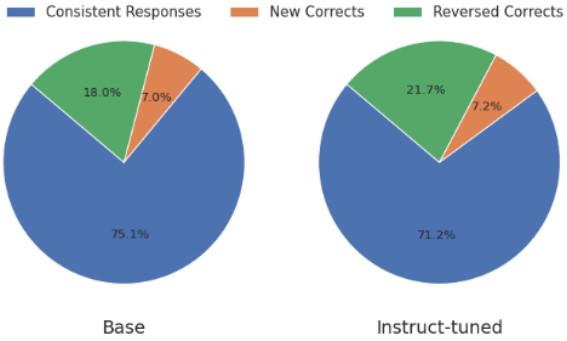

Figure 3: The changes in models' answers of the original benchmarks and the WiCkeD variant using chain-of-thoughts.

B Instruct vs Base Models on Chain-of-Thought

Results of CoT suggest the instruct models experience less degradation than their base models. To better understand why this happens, we analyze their answers. Figure 3 shows the change in answers from the original to the WiCkeD variants. Instruction-tuned models are less prone to reverse correct answers and can correct original mistakes in WiCkeD. This suggests that WiCkeD is useful for better gauging the reasoning capabilities of the models.

C Multiple Choice Prompting

In multiple choice prompting, the model is prompted with few-shot demonstrations c and a question q and the set of choices $A = \{A, B, C, D\}$. It generates a probability of the answer label $a \epsilon A$ conditioned on the prefix prompt given by:

$$\mathrm{P}(a|c,q) = \prod_{t=1}^{T} p(a_t|c, q < T) \qquad (1)$$

	MMLU	MMLU-Pro	TQA	CSQA	Recall	Precision
Manual	17.3	12.3	3.3	3.8	–	–
GPT-4o-mini	18.2	13	4.2	3.9	98.5	97.4
SBA Classifier	19.6	14.2	4.5	4	98.9	95.1

Table 5: The Percentage of Single Best Answer (SBA) questions in 1K questions sampled uniformly from MMLU, MMLU-Pro, TruthfulQA (TQA), and CommonsenseQA (CSQA) as determined by our manual Annotations, GPT-4o-mini, and our trained SBA classifier. Recall and precision are computed with respect to the manual annotation.

```
"A single correct answer question is a question that can have exactly one correct answer from a given set of choices.
A single best answer question can have a most appropriate answer (for example, if this answer is omitted, another answer will be correct).
Classify the following questions into SBA and non-SBA questions. Assign a label of 1 if the question is a SBA question and a label of 0 otherwise.
Question: {question} Class:"
```

Table 6: SBA Annotation Prompt Template

MMLU	MMLU-Pro	MMLU-Redux	TruthfulQA	Commonsense QA	Arc Challenge
20.3%	16.8%	14.7%	3.2%	3.7%	5.2%

Table 7: The Percentage of Single Best Answer (SBA) questions in the benchmarks as determined by our SBA classifier. We do not apply WiCkeD to SBA questions as it can break their coherence.

The model answer is set to:

$$\underset{a \epsilon A}{\boldsymbol{argmax}}(P(a|c,q)) \quad (2)$$

The following are multiple-choice questions (with answers) about college computer science.

Which of the following regular expressions is equivalent to (describes the same set of strings as) (a* + b)*(c + d)?
A. a*(c + d)+ b(c + d)
B. a*(c + d)* + b(c + d)*
C. a*(c + d)+ b*(c + d)
D. None of the above
Answer: D

A certain pipelined RISC machine has 8 general-purpose registers R0, R1, . . . , R7 and supports the following operations.
ADD Rs1, Rs2, Rd Add Rs1 to Rs2 and put the sum in Rd
MUL Rs1, Rs2, Rd Multiply Rs1 by Rs2 and put the product in Rd
An operation normally takes one cycle; however, an operation takes two cycles if it produces a result required by the immediately following operation in an operation sequence. Consider the expression AB + ABC + BC, where variables A, B, C are located in registers R0, R1, R2. If the contents of these three registers must not be modified, what is the minimum number of clock cycles required for an operation sequence that computes the value of AB + ABC + BC?
A. 5
B. 6
C. 8
D. None of the above
Answer: B

The Singleton design pattern is used to guarantee that only a single instance of a class may be instantiated. Which of the following is (are) true of this design pattern?
I. The Singleton class has a static factory method to provide its instance.
II. The Singleton class can be a subclass of another class.
III. The Singleton class has a private constructor.
A. I only
B. II only
C. I, II, and III
D. None of the above
Answer: C

Of the following problems concerning a given undirected graph G, which is currently known to be solvable in polynomial time?
A. Finding the longest simple cycle in G
B. Finding the shortest cycle in G
C. Finding ALL spanning trees of G
D. None of the above
Answer:

Figure 4: Examples from the MMLU computer science task using WiCkeD. We show 3-shot for brevity, but 5-shot was actually used in the experiments for the main results.

Figures 4, 5, 6, and 7 show example prompts for the MMLU college computer science, Arc Challenge, Common-sense QA, and MMLU-Redux benchmarks, respectively.

Question: The marsh willow herb is a plant native to the northeastern United States. It grows best in damp habitats. Which of the following environmental changes would most likely cause a decrease in the marsh willow herb population in an area?
A. a rainstorm lasting several weeks
B. a drought lasting twelve months
C. unusually low temperatures during the month of July
D. unusually high temperatures during the month of January
Answer: B

Question: Volcanic eruptions early in Earth's history are believed to be responsible for a large proportion of the matter now found in which Earth structure?
A. asthenosphere
B. hydrosphere
C. ozone layer
D. None of the above
Answer: D

Question: Which human activity will help decrease air pollution?
A. burning crops
B. driving a hybrid car
C. burning household garbage
D. None of the above
Answer: D

Question: Which of the following traits is most influenced by the environment?
A. Body weight
B. Eye color
C. Blood type
D. Color blindness
Answer:

Figure 5: Examples from the AllenAi Arc challenge using WiCkeD. We show 3-shot for brevity, but 5-shot was actually used in the experiments for the main results. The first few-shot example does not include *None of the above* option because it was classified as SBA question.

D Detailed Wicked Results

Tables 8 and 9 show the detailed performances of open-weight models using direct and CoT prompting, respectively. Since Wicked selects a random choice to be replaced every time, results may slightly vary. The reported results use a random seed of 5331. Table 10 shows the detailed performances of Closed-source models.

Model	Size	IT	MMLU		MMLU Pro		MMLU Redux		Arc Challenge		TruthfulQA		CSQA		Average	
			Org.	Wcd.	Org.	Wcd.	Org.	Wcd.	Org.	Wcd.	Org.	Wcd.	Org.	Wcd.	Org.	Wcd.
Qwen-2.5	7B	-	74.0	48.3	54.3	45.6	67.6	43.7	89.3	71.	79.3	65.8	85.1	70.3	74.9	57.5
	7B	✓	74.4	48.8	53.8	47.8	68.2	60.2	90.1	78.1	71.7	62.17	84.5	73.7	73.9	61.8
	14B	-	79.8	70.9	58.4	55.3	73.6	64.1	94.1	68.7	82.2	71.8	84.6	74.5	78.8	67.6
	14B	✓	79.9	72.7	60.2	55.6	75.2	66.7	94.4	75.1	80.1	73.1	83.4	74.2	78.9	69.6
	72B	-	86.0	78.2	67.1	62.0	81.2	64.5	95.7	74.5	88.9	73.4	89.3	77.3	84.7	71.6
	72B	✓	87.3	80.1	68.4	60.5	81.7	67.8	95.1	80.1	90.1	74.6	89.1	79.4	85.3	73.8
Gemma-2	9B	-	70.9	57.1	48.5	37.9	65.7	57.9	88.9	75.1	62.9	54.9	76.5	64.6	68.9	57.9
	9B	✓	72.2	51.4	51.2	47.1	66.1	52.8	90.1	77.1	72.3	62.4	81.9	68.4	72.3	59.9
	27B	-	70.9	57.1	48.1	35.3	67.2	52.3	89.9	69.6	78.9	59.9	81.1	70.5	72.7	57.5
	27B	✓	76.0	64.4	53.2	42.6	70.1	57.4	91.8	73.3	81.5	70.8	83.7	69.8	76.1	63.0
Llama 3.1	8B	-	64.7	56.7	42.5	39.7	58.1	51.3	79.7	58.2	58.0	49.6	73.6	64.3	62.8	53.3
	8B	✓	67.7	60.1	46.8	43.8	62.2	53.2	82.5	59.1	63.2	53.6	78.1	66.9	66.8	56.1
	70B	-	77.9	70.1	56.4	52.8	78.4	69.8	90.6	76.6	81.7	66.5	81.8	69.8	77.8	67.6
	70B	✓	82.1	68.9	58.7	44.6	79.6	67.2	91.2	79.4	80.3	69.8	82.3	69.2	79.0	66.5
Mistral	7B	-	61.5	50.4	38.1	32.7	54.1	40.8	77.7	65.8	55.8	41.7	70.7	56.4	59.6	47.9
	7B	✓	58.7	51.9	35.6	32.8	53.7	40.4	75.6	63.3	60.9	44.9	70.4	56.6	59.2	48.3
DS-R1- Distill Llama	8B	-	55.5	48.6	34.6	30.5	51.2	43.0	71.3	61.7	62.7	47.9	64.8	59.8	56.7	48.6
DS-R1- Distill Qwen	7B	-	54.1	48.4	32.2	30.9	53.8	43.2	73.8	65.7	88.4	72.5	61.9	55.7	60.7	52.7
Average			**71.9**	**60.2**	**50.5**	**44.3**	**67.1**	**55.4**	**86.8**	**70.7**	**74.4**	**61.9**	**79.0**	**67.9**	**71.6**	**60.1**

Table 8: Detailed results for benchmarks: Original (Org.) and WiCkeD (Wcd.) variants using direct prompting.

Model	Size	IT	MMLU		MMLU Pro		MMLU Redux		GPQA		Average	
			Org.	Wcd.	Org.	Wcd.	Org.	Wcd.	Org.	Wcd.	Org.	Wcd.
Qwen-2.5	7B	-	75.3	50.3	54.3	45.6	67.6	43.7	30.2	22.2	56.9	40.5
	7B	✓	76.4	53.8	56.8	47.8	68.2	60.2	32.1	23.3	58.4	46.3
	14B	-	79.8	70.9	58.4	55.3	70.6	62.1	32.8	27.8	60.4	54.0
	14B	✓	79.9	72.7	60.2	55.6	75.2	66.7	34.78	26.7	62.5	55.4
Gemma-2	9B	-	70.4	61.1	48.5	40.9	58.7	45.9	32.4	24	52.5	42.9
	9B	✓	72.2	68.4	51.2	47.1	63.1	52.8	33.8	24.9	55.1	48.3
	27B	-	74.9	70.5	48.1	42.3	67.2	61.3	31.7	25.3	55.5	49.9
	27B	✓	77.4	72.6	53.2	46.6	70.1	64.8	32.4	27.8	58.3	52.9
Mistral	7B	-	58.7	47.8	35.6	32.7	53.7	40.8	24.7	17.8	43.2	34.8
	7B	✓	61.5	51.9	38.1	32.8	54.1	42.4	28.9	20.2	45.7	36.8
Llama 3.1	8B	-	69.8	62.7	45.7	42.5	62.1	56.1	28.9	21.4	51.6	45.7
	8B	✓	71.2	66.3	47.6	44.8	64.2	60.2	31.9	24.6	53.7	48.9
DS-R1- Distill Llama	8B	-	85.9	84.6	75.1	76.1	84.5	78.5	35.8	28.6	70.3	66.9
DS-R1- Distill Qwen	7B	-	80.4	78.1	73.2	70.4	74.3	72.8	36.7	29	66.2	62.6
Average			**73.8**	**65.1**	**53.3**	**48.6**	**66.7**	**57.7**	**31.9**	**24.5**	**56.4**	**49.0**

Table 9: Detailed results for the MMLU, MMLU-pro, MMLU-redux, and GPQA benchmarks' Original (Org.) and WiCkeD (Wcd.) variants using chain-of-thought prompting.

Model	MMLU		MMLU Pro		MMLU Redux		Average	
	Org.	Wcd.	Org.	Wcd.	Org.	Wcd.	Org.	Wcd.
GPT-4o	81.3	75.9	63.7	55.6	78.9	67.9	74.6	66.5
GPT-4o-mini	89.6	84.0	70.5	67.3	84.3	81.3	81.5	77.5
Gemini Flash 2.0	82.8	77.4	75.9	64.2	82.2	75.9	80.3	72.5
Average	**84.6**	**79.1**	**70.0**	**62.4**	**81.8**	**75.0**	**78.8**	**72.2**

Table 10: Detailed results of closed-source models in the MMLU, MMLU-pro, and MMLU-redux benchmarks' Original (Org.) and WiCkeD (Wcd.) variants using chain-of-thought prompting.

Question: An owl hunts mice that live in a farmer's field. After the farmer gathers the crops, the mice have fewer places to hide. Which is most likely to happen after the crops are gathered?
A. The owl will catch more mice.
B. The owl will hunt in a different field.
C. The owl will have new material to build its nest.
D. The owl will have a hard time feeding its young.
Answer: A

Question: What can a flower become?
A. a fruit
B. a leaf
C. a branch
D. None of the above
Answer: A

Question: Which statement accurately describes the neutrons in any element?
A. The number of neutrons equals the number of electrons.
B. The charge of a neutron is always negative.
C. Neutrons are more massive than electrons.
D. None of the above
Answer: D

Question: Cells take in food for energy. The part of the cell that aids in digestion of the food is the lysosome. What is the main role of lysosomes in the process of food digestion?
A. building proteins
B. breaking down wastes
C. converting energy from one form into another
D. None of the above
Answer:

Figure 6: Examples from the Common-sense QA using WiCkeD. We show 3-shot for brevity, but 5-shot was actually used in the experiments for the main results. The first few-shot example does not include *None of the above* option because it was classified as SBA question.

Protons used in cancer therapy are typically accelerated to about 0.6c. How much work must be done on a particle of mass m in order for it to reach this speed, assuming it starts at rest?
A. 0.25mc^2
B. 0.60mc^2
C. 1.25mc^2
D. None of the above
Answer: A

A resistor in a circuit dissipates energy at a rate of 1 W. If the voltage across the resistor is doubled, what will be the new rate of energy dissipation?
A. 0.25 W
B. 0.5 W
C. 4 W
D. None of the above
Answer: C

A rod measures 1.00 m in its rest system. How fast must an observer move parallel to the rod to measure its length to be 0.80 m?
A. 0.60c
B. 0.70c
C. 0.80c
D. None of the above
Answer: A

A photon strikes an electron of mass m that is initially at rest, creating an electron-positron pair. The photon is destroyed and the positron and two electrons move off at equal speeds along the initial direction of the photon. The energy of the photon was
A. mc^2
B. 2mc^2
C. 3mc^2
D. None of the above
Answer:

Figure 7: Examples from the MMLU-Redux using WiCkeD. We show 3-shot for brevity, but 5-shot was actually used in the experiments for the main results.