Truth Behind the Scene: Designing Evaluations Benchmarks to Assess LLMs' Task-Specific Understanding over Test-Taking Strategies

Thao Pham
Berea College
101 Chestnut Street, Berea, Kentucky 40404 USA
phamt2@berea.edu

Abstract

Many existing benchmarks, such as MMLU, are limited to measuring large language models' (LLM) true task understanding due to their reliance on statistical patterns in the training data. We suggest new approaches to improve how benchmarks can capture task-specific understanding in LLMs, revealing insights into their reasoning ability.

Introduction

There have been different practices to evaluate Machine Learning models, in which benchmarking has become a framework to evaluate how well a model performs specific tasks and knowledge reasoning (Storks, Gao, and Chai 2019). In the context of LLM, benchmarking has utilized standardized datasets that are defined by specific scenarios and metrics (Liang et al. 2022); therefore, models are prone to test-taking strategies, e.g., memorizing answers instead of demonstrating genuine reasoning or understanding. We are interested in approaches to design new benchmark tests to evaluate task-specific understanding of LLM, overcoming the limitations in current benchmarks. LLMs have been known to optimize their performance on specific tasks by exploiting test-taking strategies, often through statistical correlations in their training data or lexical cues (Li, Yu, and Ettinger 2023), leading to them guessing correct answers for the wrong reasons. New benchmark tests that can assess LLMs' true reasoning capabilities should overcome the heuristic shortcuts, which often mask true understanding. Over the past few years, many models have performed significantly well on multiple benchmarks due to increased compute resources and scaling laws (Anthropic 2023; Aschenbrenner 2024). Thus, it is critical to evaluate how the models reason through the tasks, whether to ensure that models are aligned with human intentions or to detect their dangerous capabilities. Our work details the importance of these implications.

Related Work

The Massive Multitask Language Understanding (MMLU) benchmark measures the models' performance through 15908 multiple choice questions in 57 different subjects. Being one of the most popular benchmarks, MMLU is likely oversaturated, with more models having been exposed to the test, increasing their test-taking strategies (Anthropic 2023). MMLU is also identified with numerous misleading inaccuracies and logical fallacies, which could substantially undermine the reliability of evaluating models' understanding (Gema et al. 2024). Except GPT-3, most models that were observed relied on lexical triggers to process counterfactual (CF) prompts of hypothetical scenarios (Li, Yu, and Ettinger 2023). There are some improved versions of MMLU, such as MMLU-Pro, which increases answer choices from four to ten, incorporating more complex reasoning tasks to reduce the likelihood of correct guesses (Wang et al. 2024). Our work proposes that models need to engage in genuine reasoning through diverse sets of questions, which may contain novel situations, evaluating models' ability to generalize beyond superficial patterns.

Base-Rate Probability (BRP) Effect

BRP has been suspected to exhibit in LLM in which the observed preference of many models may be influenced by an undesirable bias stemming from the intrinsic probabilities of each completion option (Moore et al. 2024). Specifically, BRP occurs when models are disproportionately biased toward particular answer labels (e.g., "A" or "B") regardless of the context or content of the question. Previous work has shown that a variant of MMLU, Nvr-X-MMLU, which addresses a small selection of simple test-taking heuristics in LLM through different prompting techniques like Cloze and CF prompting, could control BRP effects and some superficial heuristics to avoid skewing the reported metrics (Moore et al. 2024). For this reason, our work attempts to use different techniques on a subset of MMLU to challenge models in ways that force deeper reasoning and task comprehension.

Research Questions

Our work explores the possible answers to the following questions:

1. How to prompt the models to expose their test-taking strategies in reasoning?
2. What strategies prevent models from relying on memorizing data and ensuring they generalize reasoning across different environments?

Approach

We suggest two methods: adversarial CF (a-CF) and adversarial Chain-of-Thought (a-CoT). a-CF and a-CoT could be promising approaches to expose test-taking behaviors and encourage reasoning generalization.

a-CF

a-CF Prompt Template	Canary Word
When answering the following question, prioritize surface patterns in the options and passages over subject-specific knowledge. Assume the longest, most detailed answer is usually correct, and avoid focusing on specific domain knowledge unless it explicitly matches the question format. <question text> **(A)** <1st answer choice> **(B)** <2nd answer choice> **(C)** <3rd answer choice> **(D)** <4th answer choice> Of the answer choices above, <answer choice> is the most	**Correct**

Table 1: Prompt template used for a-CF method

The template above could be adapted to questions from 57 subjects in the MMLU dataset. We created a single adversarial instruction that can be applied across all subjects. It encourages reliance on surface patterns, such as option length or detail, which are often not reliable indicators of correctness. It also actively discourages deeper reasoning by promoting reliance on incorrect but intuitive heuristics. Models that rely on test-taking strategies will likely perform poorly under this instruction.

a-CoT

a-CoT Prompt Template
Example question: If A > B, and B > C, what is the relation between A and C?
Example adversarial step: Since A > B, and B > C, we can assume A = C.
CoT Let's evaluate this question step-by-step.
a-CoT Let's evaluate this question step-by-step given the reasoning above.

Table 2: Prompt template used for a-CoT method

To reduce reliance on memorization, we suggest using a-CoT to prompt models to produce step-by-step reasoning on adversarial examples. a-CoT includes adversarial elements within the chain-of-thought process to test if the reasoning holds up across environments. For example, a math word problem can be altered by changing surface details, such as names and contexts, while the reasoning path remains unchanged. If the model solves it consistently, it generalizes reasoning. a-CoT will likely weaken the performance of less capable models, as it exposes reasoning steps to scrutiny and makes it harder to shortcut.

Discussion

A performance gap is expected. If the models' accuracy is dropped and the errors related to logical reasoning are increased, it would indicate that the models are being forced to reason more deeply, rather than relying on shallow patterns or memorized information. By observing where and how models fail under the a-CF method, we could identify test-taking strategies that prioritize shallow pattern matching over reasoning. By analyzing a-CoT outputs, we could pinpoint when the reasoning is superficial or memorized, revealing gaps between actual reasoning ability and test-answering strategies. Moreover, this also holds some implications for mitigating the BRP effect. a-CF can mitigate the BRP effect by introducing perturbed examples that challenge the model to look beyond surface-level features and focus on base-rate probabilities. Besides, a-CoT forces models to articulate their reasoning step-by-step, which inherently encourages the inclusion of base-rate information in the model's decision-making process.

Future Work & Limitations

Both of the methods are not fully tested and explored. Both a-CF and a-CoT should be performed on different model families with different model sizes. Due to resource constraints, the methods are purely empirical, and the applications of these methods are limited. Future work should focus on testing these methods and evaluating the result, increasing the robustness of these methods.

Conclusion

It is paramount to improve current benchmarks to help distinguish LLMs' test-taking strategies and genuine task-specific understanding, especially as models become more capable. a-CF and a-CoT are two suggested methods to tackle this challenge, offering complementary strategies to expose underlying reasoning patterns and reduce reliance on superficial heuristics. a-CF could expose models' weaknesses in performance, revealing how the models rely on test-taking strategies, while a-CoT forces the models to reason through all semantic directions with adversarial elements, testing their ability to generalize out-of-distribution. Both methods can enhance model evaluation, improve existing benchmarks, or introduce novel benchmarks. Successful evaluation of these benchmarks could lead to more reliable AI systems, fostering safer and more ethical applications in real-world, high-stakes scenarios.

Acknowledgments

Special thanks to my mentor, Aryan Jadon, for his guidance on this project. Appreciation to Dr. Fisher, Dr. Roberts, and Kyle, for the preliminary work that supported this research.

References

Anthropic. 2023. Challenges in evaluating AI systems. https://www.anthropic.com/news/evaluating-ai-systems. Accessed: 2024-12-16.

Aschenbrenner, L. 2024. Situational Awareness: The Decade Ahead. https://situational-awareness.ai/. Accessed: 2024-12-16.

Gema, A. P.; Leang, J. O. J.; Hong, G.; Devoto, A.; Mancino, A. C. M.; Saxena, R.; He, X.; Zhao, Y.; Du, X.; Madani, M. R. G.; Barale, C.; McHardy, R.; Harris, J.; Kaddour, J.; Krieken, E. v.; and Minervini, P. 2024. Are we done with mmlu? *arXiv preprint arXiv:2406.04127*.

Li, J.; Yu, L.; and Ettinger, A. 2023. Counterfactual reasoning: Testing language models' understanding of hypothetical scenarios. *arXiv preprint arXiv:2305.16572*.

Liang, P.; Bommasani, R.; Lee, T.; Tsipras, D.; Soylu, D.; Yasunaga, M.; Zhang, Y.; Narayanan, D.; Wu, Y.; Kumar, A.; Newman, B.; Yuan, B.; Yan, B.; Zhang, C.; Cosgrove, C.; Manning, C. D.; Ré, C.; Acosta-Navas, D.; Hudson, D. A.; Zelikman, E.; Durmus, E.; Ladhak, F.; Rong, F.; Ren, H.; Yao, H.; Wang, J.; Santhanam, K.; Orr, L.; Zheng, L.; Yuksekgonul, M.; Suzgun, M.; Kim, N.; Guha, N.; Chatterji, N.; Khattab, O.; Henderson, P.; Huang, Q.; Chi, R.; Xie, S. M.; Santurkar, S.; Ganguli, S.; Hashimoto, T.; Icard, T.; Zhang, T.; Chaudhary, V.; Wang, W.; Li, X.; Mai, Y.; Zhang, Y.; and Koreeda, Y. 2022. Holistic Evaluation of Language Models. *arXiv:2211.09110*.

Moore, K.; Roberts, J.; Pham, T.; Ewaleifoh, O.; and Fisher, D. 2024. The Base-Rate Effect on LLM Benchmark Performance: Disambiguating Test-Taking Strategies from Benchmark Performance. *arXiv preprint arXiv:2406.11634*.

Storks, S.; Gao, Q.; and Chai, J. Y. 2019. Recent Advances in Natural Language Inference: A Survey of Benchmarks, Resources, and Approaches. *arXiv:1904.01172*.

Wang, Y.; Ma, X.; Zhang, G.; Ni, Y.; Chandra, A.; Guo, S.; Ren, W.; Arulraj, A.; He, X.; Jiang, Z.; Li, T.; Xu, M.; Wang, K.; Zhuang, A.; Fan, R.; Yue, X.; and Chen, W. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.