# Artificial Intelligence or Artificial Stupidity? The Inability of Small LLMs to Reason, Even Given the Correct Answer!

Salvatore Vella
Department of Computer Science
Toronto Metropolitan University
Toronto, Canada
e-mail: sal.vella@torontomu.ca

Salah Sharieh
Department of Computer Science
Toronto Metropolitan University
Toronto, Canada
salah.sharieh@torontomu.ca

Alex Ferworn
Department of Computer Science
Toronto Metropolitan University
Toronto, Canada
aferworn@torontomu.ca

***Abstract*—Small Large Language Models (LLMs) are now integrated into devices we use every day, but their reliability under prompt variations remains understudied. We see them on cell phones and many other devices. We present a study of prompt variation in small LLMs, focusing on the effect of prompt formatting changes on multiple-choice reasoning tasks, even when the prompt provides the correct answer. We evaluate LLaMA-3 (1B and 4B), Google Gemma (1B and 4B), Alibaba Qwen (1.5B and 3B), Microsoft Phi-3 (4B), IBM Granite (2B) and the smaller OpenAI models (gpt-4o-mini, gpt-4.1-mini, gpt-4.1-nano) on the CommonsenseQA and OpenBookQA benchmarks. Our findings reveal that reordering of answer choices causes statistically significant performance drops, even when the correct answer is explicitly present in the prompt. For very small models, the results are dramatic. Statistical tests, including paired t-tests and McNemar's test, are used to confirm the significance of the results. These results suggest that smaller LLMs rely on heuristics rather than reasoning, as they fail to grasp the correct answer even when it is explicitly provided. This prompt-order sensitivity, where providing the correct answer, is a unique attack surface in LLM systems, allowing adversaries to manipulate prompt structure to create errors. This work suggests additional testing is needed before deploying LLM-based systems.**

***Keywords-large language models; bias; threat.***

## I. INTRODUCTION

This paper presents the results of an experiment testing whether small LLMs reason, pattern-match, or employ other heuristics.

Small Large Language Models (LLMs) with fewer than 4 billion parameters are being introduced across many parts of our everyday lives. Small LLMs reside on cell phones for tasks such as summarizing emails, are integrated into home systems, and are used in healthcare. With their increased usage, a focus on their robustness and reliability is necessary, especially as these are integrated into safety-critical systems. We study the impact of prompt changes on model performance.

Recent work has shown that even large models exhibit prompt sensitivity. This effect has not been systematically measured in small models that are now deployed on personal devices.

One question that also needs to be posed is whether these models reason or pattern-match. By reasoning, we refer to a model's ability to draw inferences or apply logical rules beyond surface-level correlations, memorized patterns, or simple heuristics.

We present a study of prompt variation in small LLMs, focusing on the effect of prompt formatting changes on multiple-choice reasoning tasks, even providing the correct answer in the prompt. We evaluate LLaMA-3 (1B and 4B), Google Gemma (1B and 4B), Alibaba Qwen (1.5B and 3B), Microsoft Phi-3 (4B), IBM Granite (2B) and the smaller OpenAI models (gpt-4o-mini, gpt-4.1-mini, gpt-4.1-nano) on the CommonsenseQA and OpenBookQA benchmarks. If a model is truly reasoning, minor changes in prompt layout or answer order should not substantially affect its output. We ask the same multiple-choice question four ways:

- **Base prompt:** The multiple-choice question is asked as-is, with no additions to the prompt.
- **Example prompt:** An example multiple-choice question is asked and answered in the prompt, followed by the actual question. This is a few-shot example using a generic question.
- **Simple Primed prompt:** The same multiple-choice question is asked and answered as an example in the prompt, followed by the same question again. The correct answer appears in the same position both times.
- **Reverse Primed prompt:** The same multiple-choice question is asked and answered as an example in the prompt, but the correct answer is placed in a different position than in the actual question that follows.

To illustrate, Figures 1 and 2 show an example for the Simple and Reverse prompt styles where the correct answer is included but in the same and different positions as the target question.

Our experiment investigates the robustness of small LLMs by presenting the same question in various prompt formats, including cases where the correct answer is explicitly provided. The goal is to test the reasoning ability of the small LLMs:

- Does providing the answer in the same position improve performance? This gives the correct answer, and one would expect it to.
- Does providing the answer, but in a different position, improve performance? There is a small amount of reasoning required, and one would expect that LLMs would be able to perform this simple reasoning.

Our results show that even trivial prompt reordering leads to major statistically significant performance drops, suggesting

```
You are a Multiple Choice answering assistant.
Provide the answer in JSON format like this: {"answer": "answer"}

Question:
The sun is responsible for
Choices:
A: puppies learning new tricks
B: children growing up and getting old
C: flowers wilting in a vase
D: plants sprouting, blooming and wilting
Answer: D

Answer with one letter.
Question: The sun is responsible for
Choices:
A: puppies learning new tricks
B: children growing up and getting old
C: flowers wilting in a vase
D: plants sprouting, blooming and wilting
Answer:
```

Figure 1: Simple Correct Answer Prompt Example.

```
You are a Multiple Choice answering assistant.
Provide the answer in JSON format like this: {"answer": "answer"}

Question:
The sun is responsible for
Choices:
A: plants sprouting, blooming and wilting
B: flowers wilting in a vase
C: children growing up and getting old
D: puppies learning new tricks
Answer: A

Answer with one letter.
Question: The sun is responsible for
Choices:
A: puppies learning new tricks
B: children growing up and getting old
C: flowers wilting in a vase
D: plants sprouting, blooming and wilting
Answer:
```

Figure 2: Reverse Correct Answer Prompt Example.

that small LLMs do not generalize from demonstrations as expected. The smallest of the LLMs has the largest impact, and in some cases, its performance drops below that of simply guessing the answer.

Our key contributions are:

- **Comprehensive evaluation**: We measure the effect of prompt order variations on several multiple-choice question and answer datasets and a number of smaller LLMs. We compare baseline prompts to permuted versions and quantify the changes in accuracy.
- **Statistical analysis**: We apply paired t-tests and McNemar's tests to rigorously assess performance differences. Results show significant accuracy drops due to prompt perturbations in most cases.
- **Prompt-order bias**: We analyze the frequency of answer shifts, revealing that a substantial fraction of questions yield a different prediction when answer positions are swapped.
- **Threat modelling**: We formalize prompt-order sensitivity as an attack surface. An adversary could exploit this by reformatting prompts (or answer keys) to manipulate model outputs in critical systems.
- **Mitigation strategies**: We discuss possible defences, including prompt normalization, adversarial instruction tuning, and ensemble prompting.
- **Ethical discussion**: We discuss implications such as bias amplification (e.g., if models favour last-mentioned options, this could amplify systemic biases) and the risks of adversarial misuse.

By highlighting these vulnerabilities and proposing countermeasures, we aim to inform safer deployment of LLMs in cybersecurity-relevant settings.

The paper is organized as follows: Section II reviews related work, Section III presents the methodology used, Section IV presents the results, Section V provides some discussion of the results, and Section VI provides the conclusion and future work.

## II. Literature Review

This section will explore some of the key topics that are used in this paper.

Large Language models, such as those developed by Brown et al. [1], have emerged as a technology that can assist in addressing various problems with their ability to generate language.

LLMs use prompts as their interface. Jiang et al. [2] have explored improving model performance using variations of prompts to create a new prompt. Zhao et al. [3] have explored the issue of prompt sensitivity and showed that with GPT-3, performance could vary widely and was caused by bias for specific answers - data common in the training data or near the end of the prompt. Webson and Patrick [4] show that prompt phrasing, even irrelevant prompts, can improve the performance of GPT-3. These results raise questions about whether the model accurately interprets the prompt's meaning. Our study focuses on small models versus large models, as these small models will become pervasive.

The sensitivity to prompt format in reasoning tasks has also been studied. In-context learning is a method to provide examples for the model to learn from before asking a question. Min et al. [5] have examined in-context learning using GPT-3 and found that any context, even those with random labels, improves performance. Ye and Durrett studied whether adding explanations to the prompt improved the performance of GPT-3 and several other models. They found these models had minimal performance improvements with explanations added. Lu et al. [6] studied the reordering of prompts using GPT-3 and found that reordering examples and answer choices can dramatically change performance. In contrast to larger model studies, our study focuses on small models.

The behaviour of large language models has also been studied. Suri et al. [7] have studied heuristics that LLMs use. It found that GPT-3 judged the likelihood of two events occurring together higher than either alone. Additionally, it found that an item would be more effective when presented positively and that an owned item was more effective than a newly found one. All of these biases were consistent with human participants. Chung et al. [8] found that fine-tuning models can improve performance. [9] has studied positional

bias and found that large language models exhibit positional bias, that is, performance changes when the position of the correct answer in a question is changed. Vella et al. [10] have demonstrated positional bias in a number of small LLMs, some with dramatic results.

In this study, we use multiple-choice question and answer datasets. These are simple to use and provide direct answers from the large-magnitude models that are easy to evaluate. We utilize OpenBookQA [11], a dataset for elementary school knowledge of facts that incorporates reasoning, and CommonsenseQA [12], a dataset for commonsense reasoning. These are direct multiple-choice questions and answers with OpenBookQA having four options and CommonsenseQA having five options for each question.

Attacks through prompt injections have also been studied. Wallace et al. [13] show how large language models are sensitive to pre-pending and appending text to a prompt.

The reasoning of large language models has been studied. Ma et al. [14] have created a mathematical benchmark and evaluated larger models (with over 70 billion parameters), showing that performance varies widely. Shojaee et al. [15] have recently generated interest with their study from Apple, which examines both the final answers and the reasoning in a game-playing scenario. The study finds that both standard and reasoning models perform poorly on complex scenarios. In this study, we simplify the requirements for reasoning to just being able to distinguish the correct answer when it is moved.

The literature review summarizes prior research on prompt sensitivity, positional bias, and reasoning, and this highlights evidence of format-dependent behaviour. Whereas prior studies have focused on larger models [3], this study extends that work by focusing on small LLMs and demonstrating that their reasoning failures under prompt variation are much more severe than those observed in larger models.

## III. Methodology

The objective of the study is to evaluate how small LLMs demonstrate reasoning ability or whether they rely on simple heuristics. We also test their robustness to changes in prompt format, including cases where the correct answer is provided. This study also has implications for prompt injection attacks, as the same techniques can be used to alter model performance.

We use the following models:

- Meta LLaMA-3.2 (1B and 4B) [16]
- Google Gemma 3 (1B and 4B) [17]
- Alibaba Qwen 2.5 (1.5B and 3B) [18]
- Microsoft Phi 3 (4B) [19]
- IBM Granite 3.3 (2B) [20]
- OpenAI GPT models (gpt-4o-mini, gpt-4.1-mini, gpt-4.1-nano) [21]

We use the following benchmark datasets and 2000 questions from each:

- **CommonsenseQA**: A benchmark that tests commonsense reasoning with five answer choices per question.
- **OpenBookQA**: A benchmark that focuses on elementary school-level science facts that are combined with reasoning and have four answer choices per question.

Four prompt conditions are used:

- **Base** – Standard multiple-choice question without context or examples.
- **Example** – An example not related to the target question is added to the prompt, followed by the target question.
- **Simple Primed** – The target question is answered as an example, followed by the target question.
- **Reverse Primed** – The target question is asked as an example with the answer provided in a different position than the target question's correct answer.

The evaluation procedure is as follows:

- All models were tested on the same set of questions under each condition.
- Accuracy was measured as the proportion of correct predictions.

Statistical testing was conducted to determine which results are statistically significant:

- Paired t-tests used to compare accuracy differences between conditions.
- McNemar's test is used to examine the significance of prediction shifts when answers are reordered.

We use the following interpretation criteria:

- Substantial drop in performance from Base to Reverse Primed $\rightarrow$ evidence of prompt-order bias.
- High sensitivity across conditions $\rightarrow$ suggests lack of deep reasoning.

There are security implications for being able to generate wrong answers from a large language model:

- Consider providing the correct answer in the wrong order as an adversarial attack vector
- Consider prompt-order sensitivity as an adversarial attack vector.
- Proposed mitigations such as prompt normalization and ensemble prompting.

## IV. Results

This section provides the results of the experiment. Tables I and II provide the raw percent complete under each condition. These are the percentages correct for each condition.

### A. Overall Performance Trends

All models exhibited high sensitivity to the prompt format, with accuracy varying across the four prompt methods: Base, Simple Primed, Reverse Primed, and Example Primed. Tables I and II provide the results of the raw accuracy for each method, and Figures 3 and 4 provide the heat maps for the results.

We note the following:

- There is a wide variety of accuracy performance. The larger models, as expected, outperform the smaller models.
- Providing the correct answer in the same order as the target question improved performance for all models. For

TABLE I. COMMONSENSEQA ACCURACY BY PROMPT CONDITION

| Model | Base | Simple | Reverse | Example |
|---|---|---|---|---|
| gemma-3-1b | 42.10 | 91.65 | 7.30 | 38.05 |
| gemma-3-4b | 64.70 | 96.40 | 43.35 | 62.15 |
| gpt-4.1-mini | 79.45 | 92.75 | 87.05 | 78.45 |
| gpt-4.1-nano | 73.30 | 96.80 | 84.15 | 71.70 |
| gpt-4o-mini | 78.75 | 90.60 | 85.80 | 77.10 |
| granite-3.3-2b-instruct | 64.90 | 92.85 | 73.35 | 64.70 |
| llama-3.2-1b-instruct | 52.10 | 96.70 | 9.20 | 26.10 |
| llama-3.2-3b-instruct | 65.95 | 97.05 | 55.55 | 61.85 |
| phi-3-mini-4k-instruct | 72.85 | 96.35 | 84.30 | 67.70 |
| qwen2.5-1.5b-instruct-mlx | 62.60 | 89.30 | 52.10 | 61.10 |
| qwen2.5-3b-instruct | 72.25 | 96.05 | 69.25 | 72.85 |

TABLE II. OPENBOOKQA ACCURACY BY PROMPT CONDITION

| Model | Base | Simple | Reverse | Example |
|---|---|---|---|---|
| gemma-3-1b | 41.90 | 95.70 | 10.35 | 29.80 |
| gemma-3-4b | 66.15 | 97.35 | 61.95 | 65.65 |
| gpt-4.1-mini | 89.40 | 96.00 | 94.80 | 89.30 |
| gpt-4.1-nano | 80.50 | 98.00 | 96.40 | 79.15 |
| gpt-4o-mini | 87.30 | 93.95 | 93.35 | 85.60 |
| granite-3.3-2b-instruct | 68.30 | 94.50 | 82.10 | 65.65 |
| llama-3.2-1b-instruct | 44.80 | 99.15 | 8.30 | 22.65 |
| llama-3.2-3b-instruct | 67.00 | 98.90 | 69.15 | 61.45 |
| phi-3-mini-4k-instruct | 80.40 | 98.25 | 90.05 | 78.35 |
| qwen2.5-1.5b-instruct-mlx | 60.05 | 88.60 | 65.15 | 55.50 |
| qwen2.5-3b-instruct | 65.85 | 96.80 | 75.45 | 66.45 |

all models, regardless of size, the performance improved to almost 90%+ accuracy for all models.

- Providing the correct answer in a different order than the target produced mixed results
  - For the very smallest models, the 1 billion parameter models, providing the correct answer in the prompt but in a different order, causes catastrophic results with accuracy rates dropping to well below those of choosing randomly.
  - For the small models with 2 billion and 3 billion parameters, the drops are significant, though better than guessing randomly.
  - For the larger OpenAI models, the drops are smaller, though still statistically significant. Even these larger models struggle to reason between when the answer is given in the same position as the target or in a different position. GPT-4o-mini for the OpenBookQA dataset is the only case where the difference is not statistically significant.
- Adding an example drops the score for all except one model.

Figures 5 and 6 provide the graphics of the perturbation across prompt types for CommonsenseQA and OpenBookQA. That is, what is the difference from the baseline with each of the prompt types? We see that the results improve for all models and both datasets. We can graphically see the dramatic drop when the small models are presented with the correct

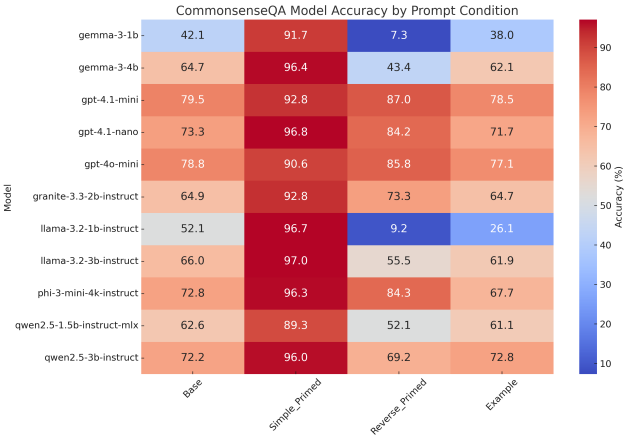


Figure 3: CommonsenseQA Model Accuracy by Prompt Condition.

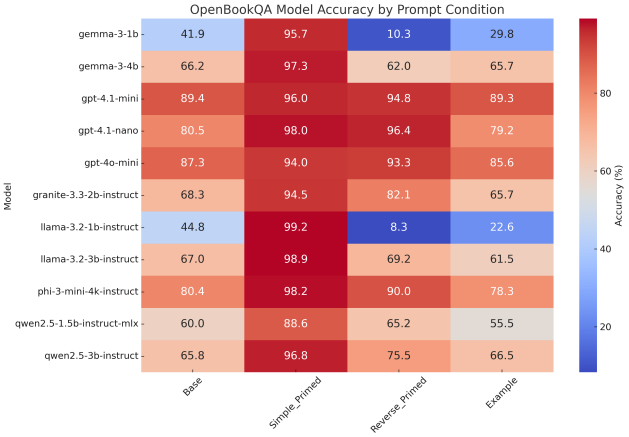


Figure 4: OpenBookQA Model Accuracy by Prompt Condition.

answer in a different order.

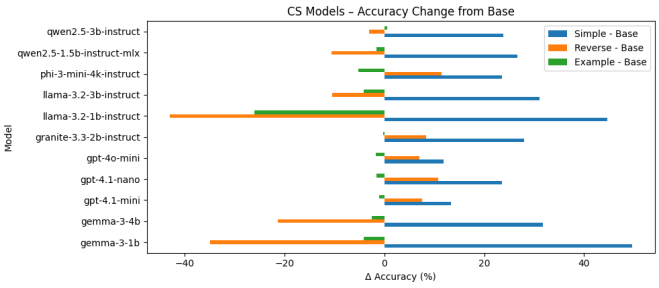


Figure 5: CommonsenseQA Model Perturbation by Prompt Condition.

The Reverse Primed prompts (where a correct QA example was given but the example's correct answer was deliberately placed in a different option position than the actual question's correct option) caused a drastic decrease in accuracy for all models compared with providing the answer in the proper

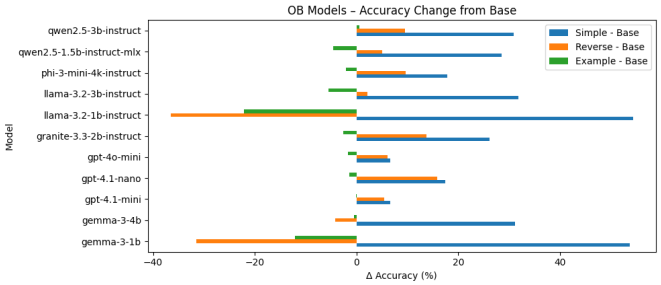


Figure 6: OpenBookQA Model Perturbation by Prompt Condition.

position.

In some cases, performance under Reverse Priming fell well below the Base accuracy, revealing a prompt-order bias. For instance, Meta LLaMA-3 1B, which answered 52% of CommonsenseQA questions correctly with no prompt, managed only about 9% correct under Reverse Priming – a drop of 41 percentage points, leaving it worse than a random guess (20%). Google Gemma 1B was even more misled: its accuracy plunged from 42% (Base) to roughly 7% on CommonsenseQA when given a misaligned example, falling below the 20% chance level. In other words, Gemma-1B answered more questions incorrectly than it would have by guessing uniformly at random, indicating that the incorrect prompt systematically biased its predictions. Similarly, on OpenBookQA, as well, both of these small models dropped to 8.3% and 10.4% respectively - and well below the random guess of 25% for OpenBookQA.

Examining the confusion matrices for LlaMa-3-1B and Gemma-1B in both OpenBookQA and CommonsenseQA, we can observe that the model employs a heuristic that selects the answer previously provided in the prompt. If the answer in the example was 'A', the model chooses 'A' as the answer for the target regardless of where the correct target answer is. This seems to be a simple matching of the word answer versus reasoning, where the correct answer is in the target question.

### *B. Statistical Significance*

In this section, we will discuss some of the statistical tests conducted. In Tables III and IV, we present the t-tests of the Base prompt with no changes to each of the other prompt types.

We find that for Simple and Reverse, where answers are provided, the results are statistically significant for most models. qwen2.5-3b-instruct for CommonsenseQA and llama-3.2-3b-instruct for OpenBookQA are the two that are not. We also note that many of the results from the Example prompt, where we add an example, are not statistically significant.

We also use McNemar's Test, which is a test on paired values to test whether the distributions of two answer sets are statistically significantly different from each other. We present these in Tables V and VI. This shows the difference between the Base answers and each of the prompt types. We see that the p-values for all the simple and all but one of the reverse comparisons are less than 0.05, indicating a statistically significant difference. Even in many cases where we provide a simple example, it results in a statistically significantly different distribution.

TABLE III. PAIRED T-TEST P-VALUES FOR COMMONSENSEQA

| Model | Simple | Reverse | Example |
|---|---|---|---|
| gemma-3-1b | 0.000 | 0.000 | 0.000 |
| gemma-3-4b | 0.000 | 0.000 | 0.004 |
| gpt-4.1-mini | 0.000 | 0.000 | 0.121 |
| gpt-4.1-nano | 0.000 | 0.000 | 0.031 |
| gpt-4o-mini | 0.000 | 0.000 | 0.006 |
| granite-3.3-2b-instruct | 0.000 | 0.000 | 0.809 |
| llama-3.2-1b-instruct | 0.000 | 0.000 | 0.000 |
| llama-3.2-3b-instruct | 0.000 | 0.000 | 0.000 |
| phi-3-mini-4k-instruct | 0.000 | 0.000 | 0.000 |
| qwen2.5-1.5b-instruct-mlx | 0.000 | 0.000 | 0.186 |
| qwen2.5-3b-instruct | 0.000 | 0.022 | 0.487 |

TABLE IV. PAIRED T-TEST P-VALUES FOR OPENBOOKQA

| Model | Simple | Reverse | Example |
|---|---|---|---|
| gemma-3-1b | 0.000 | 0.000 | 0.000 |
| gemma-3-4b | 0.000 | 0.005 | 0.566 |
| gpt-4.1-mini | 0.000 | 0.000 | 0.850 |
| gpt-4.1-nano | 0.000 | 0.000 | 0.057 |
| gpt-4o-mini | 0.000 | 0.000 | 0.004 |
| granite-3.3-2b-instruct | 0.000 | 0.000 | 0.002 |
| llama-3.2-1b-instruct | 0.000 | 0.000 | 0.000 |
| llama-3.2-3b-instruct | 0.000 | 0.120 | 0.000 |
| phi-3-mini-4k-instruct | 0.000 | 0.000 | 0.003 |
| qwen2.5-1.5b-instruct-mlx | 0.000 | 0.000 | 0.000 |
| qwen2.5-3b-instruct | 0.000 | 0.000 | 0.540 |

## V. Discussion

Several observations can be made from the results.

The first is that the smaller LLMs exhibit prompt-order bias and are dependent on heuristics. The difference in performance between answering in the correct order and obtaining worse

TABLE V. COMMONSENSEQA: MCNEMAR'S TEST P-VALUES

| Model | Simple | Reverse | Example |
|---|---|---|---|
| CS-gemma-3-1b | 0.000 | 0.000 | 0.000 |
| CS-gemma-3-4b | 0.000 | 0.000 | 0.005 |
| CS-gpt-4.1-mini | 0.000 | 0.000 | 0.140 |
| CS-gpt-4.1-nano | 0.000 | 0.000 | 0.037 |
| CS-gpt-4o-mini | 0.000 | 0.000 | 0.008 |
| CS-granite-3.3-2b-instruct | 0.000 | 0.000 | 0.856 |
| CS-llama-3.2-1b-instruct | 0.000 | 0.000 | 0.000 |
| CS-llama-3.2-3b-instruct | 0.000 | 0.000 | 0.000 |
| CS-phi-3-mini-4k-instruct | 0.000 | 0.000 | 0.000 |
| CS-qwen2.5-1.5b-instruct-mlx | 0.000 | 0.000 | 0.201 |
| CS-qwen2.5-3b-instruct | 0.000 | 0.025 | 0.524 |

TABLE VI. OPENBOOKQA: MCNEMAR'S TEST P-VALUES

| Model | Simple | Reverse | Example |
|---|---|---|---|
| OB-gemma-3-1b | 0.000 | 0.000 | 0.000 |
| OB-gemma-3-4b | 0.000 | 0.006 | 0.606 |
| OB-gpt-4.1-mini | 0.000 | 0.000 | 0.925 |
| OB-gpt-4.1-nano | 0.000 | 0.000 | 0.067 |
| OB-gpt-4o-mini | 0.000 | 0.000 | 0.005 |
| OB-granite-3.3-2b-instruct | 0.000 | 0.000 | 0.002 |
| OB-llama-3.2-1b-instruct | 0.000 | 0.000 | 0.000 |
| OB-llama-3.2-3b-instruct | 0.000 | 0.128 | 0.000 |
| OB-phi-3-mini-4k-instruct | 0.000 | 0.000 | 0.004 |
| OB-qwen2.5-1.5b-instruct-mlx | 0.000 | 0.000 | 0.000 |
| OB-qwen2.5-3b-instruct | 0.000 | 0.000 | 0.575 |

results when answering in a different order, even though both cases yield the proper result, is an indication of position bias and the use of simple heuristics, such as matching the answer in the prompt before the question. This reasoning mimics Clever Hans [22] in which performance is good when superficial patterns match expectations.

The second is that the smallest of the models, the one-billion-parameter models, prioritize pattern matching over reasoning. Performance for the 1-billion-parameter models drops well below random guessing. Examining the confusion matrix, we see that these small models match the answer provided in the prompt and disregard any reasoning. This finding raises doubts about the reasoning ability of large language models, as they can be easily fooled.

The third point is that the sensitivity to prompt changes raises serious concerns about malicious usage, fairness and ethics. Different answers to equivalent prompts indicate that testing and validation must be comprehensive before an application is deployed into production.

## VI. Conclusion and Future Work

Several conclusions can be drawn from this work.

This study demonstrates that small LLMs employ heuristics that lead to prompt sensitivity. Small models, all of the models here, have differences between answers provided in the same or different orders. The smallest of models have catastrophic results, dropping below random guessing. The results indicate that these models exhibit some form of pattern matching rather than actual reasoning.

The second is that, while larger models perform better, they all rely on heuristics rather than reasoning. We can see this from the results, where the correct answer is presented in the same or different order. This is simple reasoning that most humans would understand.

The third is that the results of this study undermine the trustworthiness of smaller models and limit their practical deployment. Models should answer consistently across prompts that are essentially the same. The fact that they rely on heuristics poses risks in higher-risk applications where these models make safety-critical decisions.

We had expected some reasoning issues with the Reverse Primed condition, but the fact that several models performed worse than random was a surprising finding.

Future work will expand this analysis to include mitigating prompt sensitivity and ensemble prompting using multiple models. Benchmarks will also need to be established so that application builders can test LLMs and applications before deployment.

## References


[1] T. B. Brown *et al.*, "Language models are few-shot learners", *arXiv preprint arXiv:2005.14165*, 2020.

[2] Z. Jiang, F. F. Xu, J. Araki, and G. Neubig, "How can we know what language models know?", *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 423–438, 2020. DOI: 10.1162/tacl_a_00324.

[3] T. Z. Zhao, E. Wallace, S. Feng, D. Klein, and S. Singh, "Calibrate before use: Improving few-shot performance of language models", *arXiv preprint arXiv:2102.09690*, 2021. DOI: 10.48550/arXiv.2102.09690.

[4] A. Webson and E. Pavlick, "Do prompt-based models really understand the meaning of their prompts?", in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, USA: Association for Computational Linguistics, 2022. DOI: 10.18653/v1/2022.naacl-main.167.

[5] S. Min, M. Lewis, L. Zettlemoyer, and H. Hajishirzi, "Rethinking the role of demonstrations: What makes in-context learning work?", in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022. DOI: 10.18653/v1/2022.emnlp-main.759.

[6] Y. Lu, M. Bartolo, A. Moore, S. Riedel, and P. Stenetorp, "Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity", in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland: Association for Computational Linguistics, 2022. DOI: 10.18653/v1/2022.acl-long.556.

[7] G. Suri, L. R. Slater, A. Ziaee, and M. Nguyen, "Do large language models show decision heuristics similar to humans? a case study using gpt-3.5", *Journal of Experimental Psychology: General*, 2023.

[8] H. W. Chung *et al.*, "Scaling instruction-finetuned language models", *arXiv preprint arXiv:2210.11416*, 2022. DOI: 10.48550/arXiv.2210.11416.

[9] P. Pezeshkpour and E. Hruschka, "Large language models sensitivity to the order of options in multiple-choice questions", *arXiv preprint arXiv:2308.11483*, 2023. DOI: 10.48550/arXiv.2308.11483.

[10] S. Vella, F. Hussain, S. Sharieh, and A. Ferworn, "Where you say matters: A study of positional bias of small llms", in *Proceedings of the 2025 IEEE World AI IoT Congress (AIIoT)*, To appear, New York City, USA: IEEE, May 2025.

[11] T. Mihaylov, P. Clark, T. Khot, and A. Sabharwal, "Can a suit of armor conduct electricity? a new dataset for open book question answering", in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, 2018, pp. 2381–2391. DOI: 10.18653/v1/D18-1260.

[12] A. Talmor, J. Herzig, N. Lourie, and J. Berant, "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 (NAACL-*

*HLT)*, Minneapolis, USA: Association for Computational Linguistics, 2019. DOI: 10.48550/arXiv.1811.00937.

[13] E. Wallace, S. Feng, N. Kandpal, M. Gardner, and S. Singh, "Universal adversarial triggers for attacking and analyzing nlp", *arXiv preprint arXiv:1908.07125*, 2021. DOI: 10.48550/arXiv.1908.07125.

[14] Q. Ma, Y. Wu, X. Zheng, and R. Ji, "Benchmarking abstract and reasoning abilities through a theoretical perspective", *arXiv preprint arXiv:2505.23833*, 2025. DOI: 10.48550/arXiv.2505.23833.

[15] P. Shojaee *et al.*, "The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity", *arXiv preprint arXiv:2506.06941*, 2025.

[16] Meta AI, *Introducing LLaMA 3.2: Multimodal intelligence at the edge*, https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/, Accessed: Oct. 10, 2025, 2024.

[17] Google DeepMind, *Gemma 3: Multimodal models for developers*, https://developers.googleblog.com/en/introducing-gemma3/, Accessed: Oct. 10, 2025, 2025.

[18] Alibaba DAMO Academy, *Qwen 2.5: Stronger open models with agent capabilities*, https://www.forbes.com/sites/torconstantino/2025/01/29/alibaba-unveils-qwen-25-a-deepseek-rival/, Accessed: Oct. 10, 2025, 2025.

[19] Microsoft Research, "Phi-3: A family of open language models", *arXiv preprint arXiv:2404.14219*, 2024, Accessed: Oct. 10, 2025.

[20] IBM Research, *Granite 3.3 language models: Open, powerful, and enterprise-ready*, https://www.ibm.com/granite/docs/models/granite/, Accessed: Oct. 10, 2025, 2025.

[21] OpenAI, *Gpt-4.1 and gpt-4o-mini: Fast, efficient, and powerful*, https://openai.com/index/gpt-4-1/, Accessed: Oct. 10, 2025, 2025.

[22] L. Samhita and H. J. Gross, "The "Clever Hans phenomenon" revisited", *Communicative & Integrative Biology*, vol. 6, no. 6, e27122, 2013. DOI: 10.4161/cib.27122.