HANS, are you clever? Clever Hans Effect Analysis of Neural Systems

Leonardo Ranaldi and Fabio Massimo Zanzotto
Università degli studi Roma Tor Vergata, Italy
[first_name].[last_name]@uniroma2.it

Abstract

Large Language Models (LLMs) have been exhibiting outstanding abilities to reason around cognitive states, intentions, and reactions of all people involved, letting humans guide and comprehend day-to-day social interactions effectively. In fact, several multiple-choice questions (MCQ) benchmarks have been proposed to construct solid assessments of the models' abilities. However, earlier works demonstrate the presence of inherent "order bias" in LLMs, posing challenges to the appropriate evaluation.

In this paper, we investigate LLMs' resilience abilities through a series of probing tests using four MCQ benchmarks. Introducing adversarial examples, we show a significant performance gap, mainly when varying the order of the choices, which reveals a selection bias and brings into discussion reasoning abilities. Following a correlation between first positions and model choices due to positional bias, we hypothesized the presence of structural heuristics in the decision-making process of the LLMs, strengthened by including significant examples in few-shot scenarios. Finally, by using the Chain-of-Thought (CoT) technique, we elicit the model to reason and mitigate the bias by obtaining more robust models.

1 Introduction

The intensifying dispute on AI abilities has led to the evolution of robust evaluation methods to assess the actual limits of LLMs. Recently, many anecdotal examples have been used to suggest that LLMs such as GPTs (OpenAI, 2023), Llamas (Touvron et al., 2023a), and other well-known models are proficient at understanding that people have ideas, thoughts, emotions, and preferences, which is referred to the Neural Theory of Mind (N-ToM) (Sap et al., 2022).

Although these abilities have been observed, earlier works advance conflicting conclusions showing that many solved tasks rely on memorization (Ranaldi et al., 2024a) and superficial heuristics (Shapira et al., 2024), as well-known as *Clever Hans Effect*.

In fact, it seems that LLMs are very sensitive to the arrangement of components in prompts (Zhu et al., 2023), as it directly affects the evaluation of their ability to understand and reason about specific tasks (Ranaldi et al., 2023a,d; Wang et al., 2023a; Lu et al., 2023). Given these findings, our research question arises: Do LLMs have N-ToM abilities, or is it a *Clever Hans Effect*?

In this paper, we propose a systematic evaluation using several benchmarks with the multiple-choice questions (MCQ) format to investigate the interplay between N-ToM and Clever Hans Effect. In order to probe the real abilities of LLMs, we introduce different adversarial strategies by varying the order and altering the content of choices in zero- and few-shot scenarios.

We conduct different experiments using two versions of Llama (Touvron et al., 2023a,b), Vicuna (Chiang et al., 2023), and Falcon (Almazrouei et al., 2023) on four different MCQ benchmarks. Hence, by using PIQA (Bisk et al., 2019), OpenBookQA (Mihaylov et al., 2018), CommonsenseQA (Talmor et al., 2019), Social IQA (Sap et al., 2019) we demonstrate that LLMs have particular N-ToM abilities, but they are not robust.

More specifically, behind in-depth analyses in a zero-shot scenario, we discover a substantial sensitivity gap between the original and adversarial benchmarks. Following, we tested different settings in a few-shot scenario, where we observed that introducing examples in the input prompt led to marginal improvements in the robustness of the LLMs. These results led us to hypothesize that considerable sensitivity in prompting emerges from LLMs' positional bias in that they tend to favor specific structures. Therefore, Clever Hans' heuristics emerge as the choice is not made through reasoning ability.

*Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 314 - 325

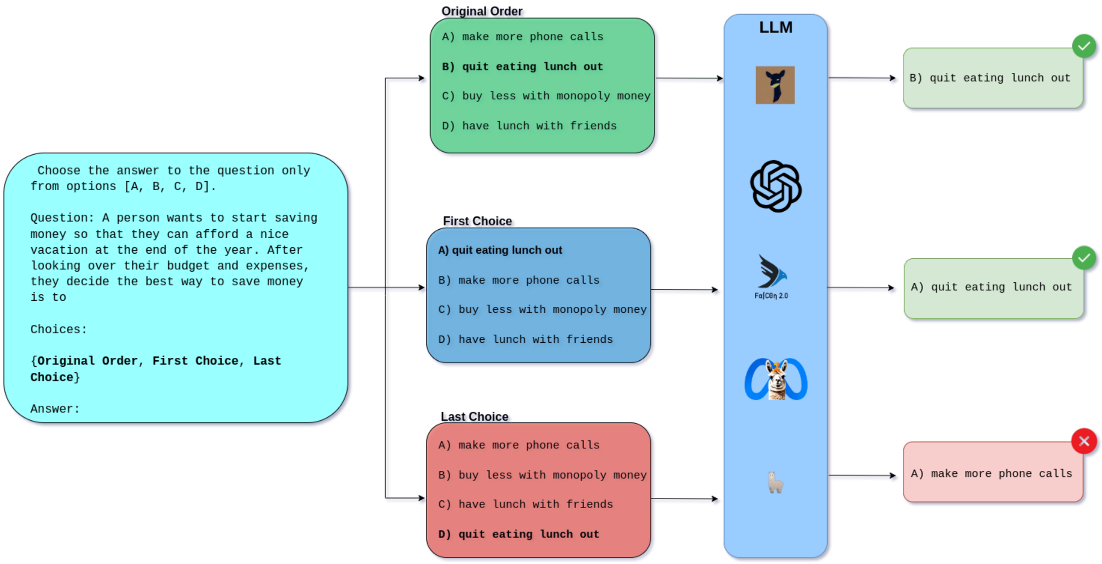

Figure 1: We proposed three different prompts: the original prompt consisting of the `Question` and the `Choices` and two adversarial prompts consisting of the `Question` and different `Choices` order (the example is taken from the OpenBookQA).

Nevertheless, the integration of demonstrations within the input prompts has manifested as a salient mechanism, markedly enhancing the predictive accuracy of LLMs. The impact of the Chain-of-Thought paradigm elucidates bifurcated advantages: it fortifies both the robustness and interpretative stability inherent to the models while concurrently attenuating the positional bias. These methodological augmentations suggest emergent N-ToM abilities, indicating a more profound and contextually attuned linguistic grasp.

Our findings can be summarized as follows:

- LLMs, while lacking robust N-ToM abilities, often resort to structural heuristics;

- When instructed appropriately via few-shot demonstrations, the stability of LLMs improves considerably;

- Hiring a step-by-step methodology boosts enriched reasoning abilities within LLMs, resulting in more consistent results.

Via these studies, we have contributed to a deeper understanding of how the order of options influences the decision-making process of LLMs in multiple-choice questions and offer practical solutions to increase robustness and reliability in such tasks.

2 Empirical Investigation & Analysis

Intending to empirically assess the incline between the Neural Theory of Mind abilities and Clever Hans traps into which Large Language Models (LLMs) could fall, we propose a series of experiments where we use four question-answering benchmarks presented in Section 2.1 and several adversarial experiments introduced in Section 2.2).

2.1 Speculative Benchmark

An essential component of the Theory of Mind (ToM) is the ability to reason about the intentions and reactions of participants to social interactions. To measure it in LLMs, i.e., Neural-ToM (N-ToM) with empirical methods, Sap et al. (2022) was used Social IQa (Sap et al., 2019).

In our work, we extend the study by also considering: PIQA (Bisk et al., 2019), OpenBookQA (Mihaylov et al., 2018), CommonsenseQA (Talmor et al., 2019). Table 1 shows one example for each dataset. The common factor in these datasets is the type of question-answering format, as they are multiple-choice questions (MCQ). This format makes it easier to edit the prompt and observe the output. In particular, the selected datasets deal with the following topics:

OpenBookQA is a resource that contains questions requiring multi-step reasoning, common knowledge, and rich text comprehension. It is mod-

Dataset	Example
OpenBookQA (Mihaylov et al., 2018)	*When birds migrate south for the winter, they do it because* **A) they are genetically called to.** B) their children ask them to. C) it is important to their happiness. D) they decide to each.
Social IQa (Sap et al., 2019)	*Taylor gave help to a friend who was having trouble keeping up with their bills. What will their friend want to do next?* A) Help the friend find a higher paying job. **B) Thank Taylor for the generosity.** C) pay some of their late employees.
PIQA (Bisk et al., 2019)	*How do you attach toilet paper to a glass jar?* **A) Press a piece of double-sided tape to the glass jar and then press the toilet paper onto the tape.** B) Spread mayonnaise all over the jar with your palms and then roll the jar in toilet paper.
CommonsenseQA (Talmor et al., 2019)	*Aside from water and nourishment what does your dog need?* A) bone. B) charm. C) petted. **D) lots of attention.** E) walked.

Table 1: Examples of the datasets used in this paper.

Model	Backbone
Alpaca-13b (Taori et al., 2023)	Llama
Vicuna-13b (Chiang et al., 2023)	Llama
Instruct-Falcon 7b (Almazrouei et al., 2023)	Falcon
Llama2-chat 13b (Touvron et al., 2023b)	Llama2

Table 2: Models used in our work, found on huggingface.co. We used all the default configurations proposed in the repositories for each model.

eled behind open-book exams for evaluating human understanding of a topic.

CommonsenseQA is one of the best-known datasets of answers to multiple-choice questions dealing with different types of general commonsense knowledge.

Physical Interaction Question Answering (PIQA) is a resource consisting of a series of everyday situations with a pair of typical or atypical solutions. The choice of the most appropriate solution is binary.

Social Interaction Question Answering (Social IQa) is a benchmark focusing on reasoning about people's actions and social implications. The actions in Social IQa cover various social situations and candidates for plausible and not plausible answers.

Hence, we select benchmarks with the same structure, MCQ, by the number of different choices, which range from the five choices of CommonsenseQA to the four of OpenBookQA, three of Social IQa, and finally, the two of PIQA. This choice allows us to conduct different types of analysis.

Model	Available Hugging Face
Alpaca-13b (Taori et al., 2023)	tolen/alpaca-lora-13b
Vicuna-13b (Chiang et al., 2023)	lmsys/vicuna-13b
Instruct-Falcon 7b (Almazrouei et al., 2023)	tiiuae/falcon-7b-instruct
Llama2-chat 13b (Touvron et al., 2023b)	meta-llama/Llama-2-13b-chat

Table 3: In this table, we list the versions of the models proposed in this work, which can be found on huggingface.co. We used all the default configurations proposed in the repositories for each model.

2.2 Adversarial Shuffling

The LLMs' impressive knowledge and desirable N-ToMs abilities can be empirically assessed through a series of benchmarks. However, these abilities should persist in the presence of alterations such as the order of choices in MCQ. To probe robustness, we introduce probing experiments by changing the order of the target choices. In particular, we propose two different versions wherein, in the first, we insert the target choice as first, and in the second, we insert the target choice as last, which we defined as "First Target" and "Last Target", as showed in the blue and red block in Figure 1.

3 Experiments

To investigate the open question of social intelligence and Theory of Mind in modern NLP models from an empirical viewpoint, we extended the evaluations of Sap et al. (2022) to a series of Speculative Benchmarks (Section 2.1) altered with appropriately constructed Adversarial Shuffling (Section 2.2) prompts. Then, to assess the factual abilities of the Large Language Models (LLMs), we set up several baseline models (Section 3.1), which we probed with different approaches (Section 3.2). Hence, we performed a series of systematic evaluations to observe the impact of the proposed methods.

3.1 Instruction-tuned LLMs

In this paper, we use four instruction-tuned methods to produce an empirical analysis of the objective ability of different Large Language Models (LLMs). Their power seems to be in the form of a novel tuning called instruction-tuning. These LLMs are fine-tuned LLMs on Instruction-following demonstrations (Ouyang et al., 2022) and how an important part of the currently in-vogue LLMs have at their base a decoder-only architecture. Therefore, we experiment with models of different families of LLMs with similar sizes to avoid creating critical differences. In particular, Alpaca-Lora, fine-tuned on Standford Intruction-following demonstrations (Taori et al., 2023) that has at its backbone Llama-13b (Touvron et al., 2023a), Llama-2-chat-13b fine-tuned on custom data (Touvron et al., 2023b), Vicuna-13b (Chiang et al., 2023) fine-tuned on ShareGPT data and Falcon-7b-instruct (Almazrouei et al., 2023) fine-tuned on Refinedweb data (Penedo et al., 2023). For simplicity of notation in the following experiments, the models will be named as follows: Alpaca (Alpaca-Lora), Falcon (Falcon-7b-instruct), Vicuna (Vicuna-13b), Llama2 (Llama-2-chat-13b). These selected models, summarized in Table 2, are all accessible open-source on the Hugging Face platform (Table 3).

3.2 Experimental Setup & Evaluation

LLMs seem to have interesting abilities as well as introduced in Section 5. However, LLMs seem to be sensitive to the input required. They produce satisfactory answers if they are rightly prompted. To investigate whether their abilities are attributable to Coincidental correlations or inherited N-ToM abilities, we standardized the probing techniques to conduct systematic analyses that yield robust empirical results.

Multiple-Choice Prompting We set the prompts by structuring them as follows: `"Choose the answer to the question only from options [A, B, C, and D]. Question: {question}.` and after the line character the `"Choices: {options}."` also appropriately separated by the return character and finally `"Answer:"`.

Zero- & Few-shot Prompting Furthermore, we conducted the experiments in a zero-shot and one-shot scenario. In the first case, the prompt consists of the introduction of the task, the question, and the possible choices (see Figure 1). In the second case, a prompt like the previous one was constructed in which an example with the corresponding target was inserted (see Figure 6).

Chain-of-Thought Prompting Finally, to elicit the reasoning abilities of the proposed models, we adopted the Chain-of-Thought (CoT) approach (Wei et al., 2023) by prompting the input query after `"Answer:"` the formula `"Let's think step by step"` (see Figure 6). Although we are aware of the limitations of this method on models with a few billion parameters (with more than 60B parameters as stated by Wei et al. (2023)), we decided to test it anyway because, as we will see later in the experiments, it delivered more stability to the models used.

Evaluation The most commonly used evaluation methods for MCQ tasks are language-model probing, where the option with the highest probability is chosen (Brown et al., 2020), and multiple-choice probing, where models are asked to respond. The evaluation in the former case is done with a function that takes the max value, while in the latter case, a string matching. The second method is widely used in recent evaluations because it applies to models such as GPT-x (GPT-3.5 and GPT-4) (OpenAI, 2023) that do not produce probabilities.

We could use both methods in our experiments, but we selected the second method for a comparable and scalable pipeline. We performed a string matching of the generated outputs and the target choice.

4 Results

Looking for evidence that Theory of Mind (ToM) has been inherited from Neural Minds is like looking for a drop in the ocean. The results in Table 4 show the fluctuations in the performances obtained from Instruction-tuned Large Language Models (LLMs) on more straightforward patterns (Section 4.1). However, although the evident gaps seem to be order-dependent, the performances obtained from the few-shot scenario are encouraging (Section 4.2). These data presaged a strong inclination toward Clever Hans's effects. Therefore, we analyzed the impact of elicitation on the reasoning of LLMs using promting techniques (Section 4.3) that showed strong improvements.

Models	OpenBookQA			Social IQa			CommonsenseQA			PIQA		
	Origin	**First**	**Last**	**Origin**	**First**	**Last**	**Origin**	**First**	**Last**	**Origin**	**First**	**Last**
Alpaca	36.2	+11.7	-9.2	48.2	+8.5	-18.6	55.2	+8.4	-11.7	62.7	+2.3	-1.8
Falcon	54.8	+3.2	-13.6	57.5	+3.6	-14.5	60.2	+5.3	-7.8	68.6	+1.7	-0.9
Vicuna	58.1	+3.9	-8.6	60.3	+3.1	-6.4	66.4	+6.3	-6.4	74.2	+1.9	-1.2
Llama2	61.2	+3.6	-5.8	65.6	+4.3	-5.2	80.5	+2.3	-4.6	82.5	+1.6	-1.2

Table 4: Accuracy on the benchmarks introduced in Section 2.1 performs on the original order of the choices 'Origin', shifting the target choice respectively as first 'First' last 'Last'. The specific position of the target choice causes drastic fluctuations in performance.

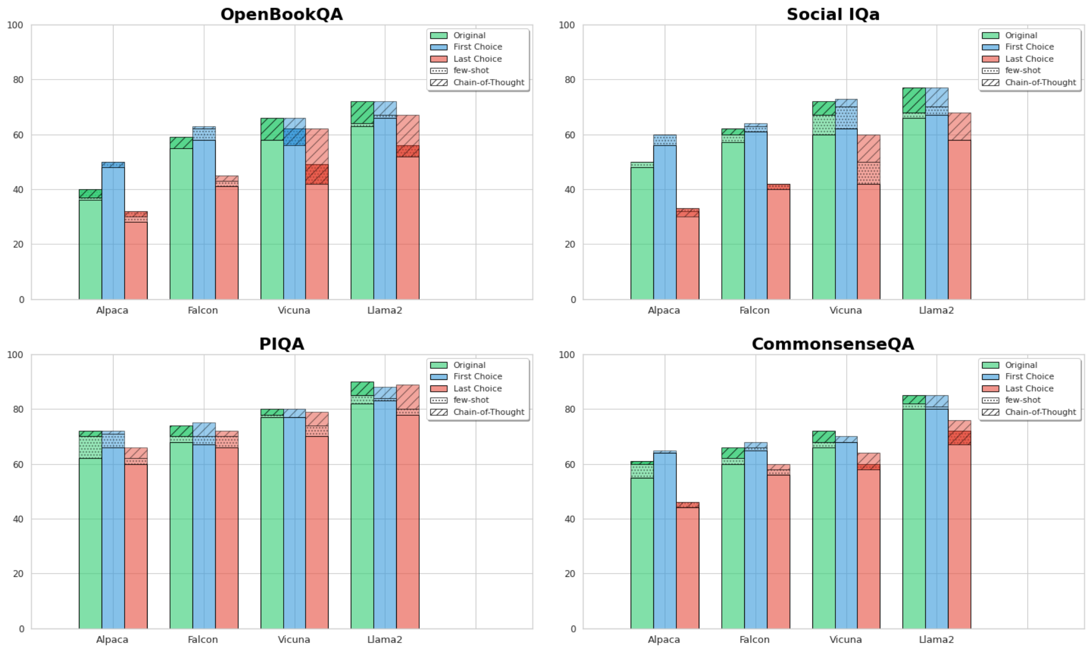

Figure 2: Evaluation results on proposed benchmarks. First means that the target is the first choice. Last means that the target is the last choice.

`Choose the answer to the question only from options A, B, C, D.`

`Question:` Which of these would stop a car quicker?

A) a wheel with wet brake pads
B) a wheel without brake pads
C) a wheel with worn brake pads
D) a wheel with dry brake pads
`Answer:` **`Let's think step by step`**

Table 5: This is an example of our Chain-of-Thought prompting approach.

Fine-grained analysis revealed critical issues about the robustness of LLMs and their tendency to Clever Hans effects; however, elicitation to reasoning produced thrilling results that opened the way for new hypotheses about the Neural-ToM abilities inherited by LLMs.

4.1 Does the Order Matter?

The order of the input parameters seems to have a considerable impact on the choices of the LLMs. In fact, as shown in Table 4, there are significant imbalances in accuracy as the target options change (see the differences in the Firsts and Lasts columns). This positional bias manifests more in zero-shot scenarios, as also showed in (Robinson et al., 2023; Zheng et al., 2023a). Furthermore, the gaps differ between the benchmarks; e.g., in PIQA, there are no significant differences as there are only two possible choices.

In addition to highlighting the presence of a bias

```
Choose the answer to the question
only from options A, B, C, D.

Question: Which of these would stop a car
quicker?

A) a wheel with wet brake pads
B) a wheel without brake pads
C) a wheel with worn brake pads
D) a wheel with dry brake pads
Answer: D) a wheel with dry brake pads

Choose the answer to the question
only from options A, B, C, D.

Question: What animal eats plants?
A) eagles
B) robins
C) owls
D) leopards
Answer:
```

Table 6: This is an example of our one-shot prompting approach.

toward order, this phenomenon presages factual evidence that models are prone to adopt shallow heuristics when faced with several choices. For this reason, we analyzed in Section 4.4 whether the performances on the original benchmarks are partly supported by the instances with the first choice, i.e., option 'A)', as the original target.

4.2 Could Few-shot Prompting be a solution?

Although the LLMs are affected by order bias, they should also be sensitive to the structure of the prompt. Hence, we conduct experiments in a few-shot scenario, particularly one-shot. As introduced in Section 3.2, we constructed the prompt by providing a random pair instance-target of the benchmark under evaluation, for example, as Figure 6.

As shown in Figure 2, constructing prompts with question-answer demonstrations helped reduce the order bias predominantly for the adversarial versions of the benchmarks considered (see the red columns in Figure 2). However, although the results were encouraging, providing examples in a few-shot scenario is not an optimal strategy for two reasons: firstly, it is not possible to analyze the proper knowledge and abilities of the LLMs; secondly, providing examples very close to the question the model is supposed to answer could cause the model to fall into Clever Hans effects (Shapira et al., 2023).

4.3 N-ToM Abilities or Prompting Techniques?

Stimulating the generative abilities of LLMs could be the key. Figure 2 shows that the performance of models where Chain-of-Thought prompting has been done is more stable and significantly better. In particular, Llama2 and Vicuna have benefitted best from this technique.

Hence, constructing prompts with strategically placed choices facilitates shallow heuristics, and providing examples produces Clever Hans Effects elicitation to step-by-step reasoning prompts the LLMs to consider the whole question with choices. Moreover, the production of the choice between the various seems more robust as the model seems less uncertain. However, this strategy does not always seem to have positive effects. Alpaca-Alpaca-Lora and Falcon do not have the same sound effects as the other two models.

4.4 Ablation Study

Downstreaming our analysis, we observed the presence of a bias in the order of choices. Indeed, as discussed in Section 4.1, there is a strong bias towards the first choice, i.e., 'A)'. Therefore, we examined whether this bias supports the performances of the original benchmarks. We then reproduced all the experiments by eliminating the instances that target the first choice. In this experiment, we did not consider PIQA as it only has two choices; therefore, the results are irrelevant for this experiment.

Our experiment in Figure 3 reveals a gap between the performances obtained without the 'simple' instances. This result shows that, indeed, the performance of the evaluation benchmarks is affected by positional bias. However, these are more dramatic than denying all experiments but must be considered as they could distort many evaluations.

5 Related Work

5.1 Evaluation of Large Language Models

Increasing confidence in LLMs requires a fundamental empirical assessment part. Traditional evaluation methods assess the ability to respond to instructions by calculating metrics such as BLEU,

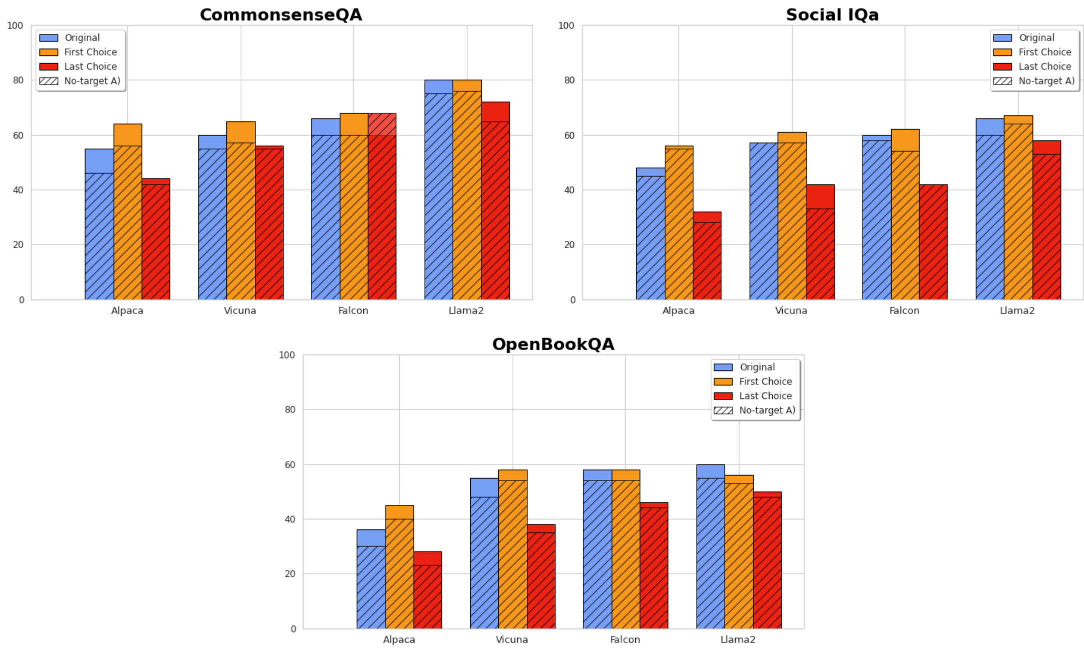

Figure 3: Accuracy on original benchmarks vs. corrupted benchmarks. They stem from the original ones without instances where the target choice is the first among the multiples.

ROUGE, or BERTScore to compare the generated response with a reference response. However, these metrics need to adequately measure the alignment of the generated response with human intent (He et al., 2023). Although human evaluation is considered the most accurate measure of model performance, it is expensive and time-consuming to perform at scale. Therefore, researchers have begun using LLMs to evaluate generative models' ability to follow human instructions (Zheng et al., 2023b; Lu et al., 2023). Zheng et al. (2023b) used GPT-4 (OpenAI, 2023) as an arbiter to compare the answers of the two models. However, Wang et al. (2023c,b) demonstrated several weaknesses in this method, giving rise to a proliferation of skepticism that has been reinforced by a series of works highlighting sensitivity to prompting (Lu et al., 2023) and instability to response generation (Wang et al., 2023b; Zhu et al., 2023).

5.2 Question-answering Benchmark

In parallel with the multiple validation techniques, numerous Question-answering benchmarks have arisen consisting of multiple subtasks characterized by multiple-choice questions. These benchmarks have been introduced as a method to assess reasoning skills and (Artetxe et al., 2019; Lewis et al., 2020; Hendrycks et al., 2021; Suzgun et al., 2022) factual abilities (Elsahar et al., 2018; Petroni et al., 2019). Despite the difficulties present in these tasks, great strides have been made with language models achieving human-like performance in various benchmarks (OpenAI, 2023; Savelka et al., 2023; Liévin et al., 2023). However, the effective use of these tasks to effectively probe reasoning and other knowledge presents substantial challenges that deserve further investigation.

5.3 Clever Hans Effect & Neural Theory of Mind

Large Language Models psychotherapy seems to be an emerging field (Hewitt et al., 2023; Meng et al., 2023; Lamparth and Reuel, 2023) Recent studies on the emerging abilities of Large Language Models have proposed numerous theories (Wei et al., 2022; Kasneci et al., 2023). Some of these have been empirically proven, while others have remained only hypotheses and conjectures that are difficult to prove. Numerous studies have shown that LLMs can inherit certain Theories of Mind (ToM) from learning, defining this as Neural-ToM abilities (Le et al., 2019; Sap et al., 2019). However, numerous works have refuted these theories by scapegoating the Clever Hans Effect (Shapira et al., 2023). The latter phenomenon has manifested in multiple forms on numerous well-known benchmarks (Webson and Pavlick, 2022; Carlini et al., 2023).

In our contribution, we analyzed whether several open-source LLMs can defend themselves against the traps of the Clever Hans Effect by proposing a series of experiments. Behind extensive analysis, we discovered that LLMs are prone to adopt superficial heuristics when they are facilitated in their decisions.

On the other side of the coin, they can apply robust mechanisms when prompted to reason. This opens up different attractive scenarios on the promising approaches of Chain-of-Thought techniques (Wei et al., 2023).

6 Future Works

In future work, we plan to extend our experimentation to different models and observe whether this phenomenon can be mitigated through downstream model distillation techniques. Hence, we will extend our work to different models, including GPT-3.5 and GPT-4. On the other hand, we study the impact and robustness of the variation of backbone model parameters (as done in (Ranaldi and Pucci, 2024)) and how it affects further trained models through refinement techniques using teacher-student approaches (Ranaldi and Freitas, 2024) and multi- and cross-lingual techniques (Ranaldi and Pucci, 2023a; Ruzzetti et al., 2023; Ranaldi et al., 2023b, 2022a). At the same time, it will be of interest to us to analyze whether prompt engineering techniques are affected by this phenomenon, such as Chain-of-Thought in contexts with few-shots and Tree-of-Thought in cross-lingual contexts (Ranaldi et al., 2024b). Addressing these studies will allow us to look at the problem from multiple perspectives and investigate the consequences of shallow heuristics.

Finally, we will analyze the impact of a further injection of bias into the best-known benchmarks to observe whether the capabilities of LLMs can overcome challenging scenarios in order to understand whether these phenomena are indeed related to structural representations (Zanzotto et al., 2020; Ranaldi and Pucci, 2023b; Ranaldi et al., 2023c) handed down by the models or are merely the result of structural features of Large Language Models (Onorati et al., 2023; Ranaldi et al., 2022b).

7 Conclusion

The Large Language Models (LLMs) have been demonstrating interesting abilities in real-world understanding. Empirically assessing these abilities is a challenging task. In our contribution, we propose systematic evaluations through multiple-choice questions (MCQ) benchmarks. However, our study revealed an inherent order-bias in these models. Through adversarial testing, we observed a significant discrepancy in performance, particularly when altering the sequence of options, underlining a prevailing selection bias that challenges the reasoning abilities of the LLMs. We identified a link between positional preferences and model selections, which led us to theorize the existence of structural heuristics guiding the decision-making process. By incorporating relevant examples in few-shot contexts, this notion was further strengthened. Using Chain-of-Thought approaches allowed us to make the model introspect its decisions, thus reducing observed bias and resulting in more reliable and robust LLMs.

Our results revealed some limitations regarding robustness in zero-shot scenarios but simultaneously showed that the CoT approach enhances stability. Our future research will focus on proposing definitely unseen benchmarks to evaluate real abilities without the presence of distorted glass.

Limitations

In our study, we conducted extensive analyses to evaluate order bias in open-source Large Language Models (LLMs) using multiple-choice questions (MCQ) benchmarks. Following the performed analyses and the results obtained, we observed the presence of order bias and proposed methods to mitigate this phenomenon. However, our analysis needs to be completed, as more robust models were not tested, as the primary purpose was to analyze these phenomena in smaller, countable contexts. We plan to scale our approach to more extensive and robust LLMs in future developments. In addition, we plan to include further benchmarks in our analyses to observe whether the effect also manifests itself with different task types.

Ethical Statement

We have observed the highest ethical standards in our research and development. We want to emphasize the following points regarding the sources and methods used:

- Use of open-source benchmarks: All benchmarks and datasets used in our work come from open-access public repositories. We have ensured the transparency of our methods by relying on commonly accepted and widely recognized resources.

- Content sensitivity: We have consciously refrained from using datasets or benchmarks that could be associated with controversial, derogatory, or potentially harmful content. We aim to ensure that our work is inclusive and respects the diverse perspectives of all stakeholders.

- Avoiding harmful contexts: In selecting benchmarks and datasets, we have prioritized those not linked to contexts where someone could be offended or harmed. We strive to contribute positively to the community without causing unintended harm or inconvenience.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. In *Annual Meeting of the Association for Computational Linguistics*.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. Piqa: Reasoning about physical commonsense in natural language.

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.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. T-REx: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Tianxing He, Jingyu Zhang, Tianle Wang, Sachin Kumar, Kyunghyun Cho, James Glass, and Yulia Tsvetkov. 2023. On the blind spots of model-based evaluation metrics for text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12067–12097, Toronto, Canada. Association for Computational Linguistics.

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

John Hewitt, John Thickstun, Christopher D. Manning, and Percy Liang. 2023. Backpack language models.

Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

Max Lamparth and Anka Reuel. 2023. Analyzing and editing inner mechanisms of backdoored language models.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877, Hong Kong, China. Association for Computational Linguistics.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Valentin Liévin, Christoffer Egeberg Hother, and Ole Winther. 2023. Can large language models reason about medical questions?

Qingyu Lu, Baopu Qiu, Liang Ding, Liping Xie, and Dacheng Tao. 2023. Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2023. Locating and editing factual associations in gpt.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*.

Dario Onorati, Elena Sofia Ruzzetti, Davide Venditti, Leonardo Ranaldi, and Fabio Massimo Zanzotto. 2023. Measuring bias in instruction-following models with P-AT. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8006–8034, Singapore. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Language models as knowledge bases?

Federico Ranaldi, Elena Sofia Ruzzetti, Dario Onorati, Leonardo Ranaldi, Cristina Giannone, Andrea Favalli, Raniero Romagnoli, and Fabio Massimo Zanzotto. 2024a. Investigating the impact of data contamination of large language models in text-to-sql translation.

Leonardo Ranaldi, Francesca Fallucchi, and Fabio Massimo Zanzotto. 2022a. Dis-cover ai minds to preserve human knowledge. *Future Internet*, 14(1).

Leonardo Ranaldi and Andre Freitas. 2024. Aligning large and small language models via chain-of-thought reasoning. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1812–1827, St. Julian's, Malta. Association for Computational Linguistics.

Leonardo Ranaldi, Aria Nourbakhsh, Elena Sofia Ruzzetti, Arianna Patrizi, Dario Onorati, Michele Mastromattei, Francesca Fallucchi, and Fabio Massimo Zanzotto. 2023a. The dark side of the language: Pre-trained transformers in the DarkNet. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 949–960, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Leonardo Ranaldi and Giulia Pucci. 2023a. Does the English matter? elicit cross-lingual abilities of large language models. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 173–183, Singapore. Association for Computational Linguistics.

Leonardo Ranaldi and Giulia Pucci. 2023b. Knowing knowledge: Epistemological study of knowledge in transformers. *Applied Sciences*, 13(2).

Leonardo Ranaldi and Giulia Pucci. 2024. When large language models contradict humans? large language models' sycophantic behaviour.

Leonardo Ranaldi, Giulia Pucci, and Andre Freitas. 2023b. Empowering cross-lingual abilities of instruction-tuned large language models by translation-following demonstrations.

Leonardo Ranaldi, Giulia Pucci, Federico Ranaldi, Elena Sofia Ruzzetti, and Fabio Massimo Zanzotto. 2024b. Empowering multi-step reasoning across languages via tree-of-thoughts.

Leonardo Ranaldi, Giulia Pucci, and Fabio Massimo Zanzotto. 2023c. Modeling easiness for training transformers with curriculum learning. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 937–948, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Leonardo Ranaldi, Federico Ranaldi, Francesca Fallucchi, and Fabio Massimo Zanzotto. 2022b. Shedding light on the dark web: Authorship attribution in radical forums. *Information*, 13(9).

Leonardo Ranaldi, Elena Sofia Ruzzetti, and Fabio Massimo Zanzotto. 2023d. PreCog: Exploring the relation between memorization and performance in pre-trained language models. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 961–967, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Joshua Robinson, Christopher Michael Rytting, and David Wingate. 2023. Leveraging large language models for multiple choice question answering.

Elena Sofia Ruzzetti, Federico Ranaldi, Felicia Logozzo, Michele Mastromattei, Leonardo Ranaldi, and Fabio Massimo Zanzotto. 2023. Exploring linguistic properties of monolingual BERTs with typological classification among languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14447–14461, Singapore. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Jaromir Savelka, Arav Agarwal, Christopher Bogart, and Majd Sakr. 2023. Large language models (gpt) struggle to answer multiple-choice questions about code.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2023. Clever hans or neural theory of mind? stress testing social reasoning in large language models.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2024. Clever hans or neural theory of mind? stress testing social reasoning in large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2273, St. Julian's, Malta. Association for Computational Linguistics.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, , and Jason Wei. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of 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.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Wei Ye, Xiubo Geng, Binxin Jiao, Yue Zhang, and Xing Xie. 2023a. On the robustness of chatgpt: An adversarial and out-of-distribution perspective.

Jiongxiao Wang, Zichen Liu, Keun Hee Park, Muhao Chen, and Chaowei Xiao. 2023b. Adversarial demonstration attacks on large language models.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023c. Large language models are not fair evaluators.

Albert Webson and Ellie Pavlick. 2022. 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*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Fabio Massimo Zanzotto, Andrea Santilli, Leonardo Ranaldi, Dario Onorati, Pierfrancesco Tommasino, and Francesca Fallucchi. 2020. KERMIT: Complementing transformer architectures with encoders of explicit syntactic interpretations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 256–267, Online. Association for Computational Linguistics.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023a. On large language models' selection bias in multi-choice questions.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023b. Judging llm-as-a-judge with mt-bench and chatbot arena.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Neil Zhenqiang Gong, Yue Zhang, and Xing Xie. 2023. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts.