Investigating causal understanding in LLMs

Marius Hobbhahn*
University of Tübingen

Tom Lieberum
Independent

David Seiler
Independent

Abstract

We investigate the quality of causal world models of LLMs in very simple settings. We test whether LLMs can identify cause and effect in natural language settings (taken from BigBench) such as "My car got dirty. I washed the car. Question: Which sentence is the cause of the other?" and in multiple other toy settings. We probe the LLM's world model by changing the presentation of the prompt while keeping the meaning constant, e.g. by changing the order of the sentences or asking the opposite question. Additionally, we test if the model can be "tricked" into giving wrong answers when we present the shot in a different pattern than the prompt. We have three findings. Firstly, larger models yield better results. Secondly, k-shot outperforms one-shot and one-shot outperforms zero-shot in standard conditions. Thirdly, LLMs perform worse in conditions where form and content differ. We conclude that the form of the presentation matters for LLM predictions or, in other words, that LLMs don't solely base their predictions on content. Finally, we detail some of the implications this research has on AI safety.

1 Introduction

We think that the quality of causal world models of LLMs matters for AI safety and alignment. More capable LLMs are supposed to assist humans in important decisions and solve scientific questions in the future. For all of these applications, it is important that the LLM's causal world model is accurate.

We can see many failure modes stemming from inaccurate causal world models. On a small scale, an LLM that is used as a personal assistant might give plausibly sounding but ultimately incorrect advice to its users. On a larger scale, LLMs might be used to conduct or assist with important scientific questions. In case their causal world models are false, the results from their scientific predictions are likely wrong and could harm people affected by the consequences. On an even larger scale, a very powerful LLM that is able to take actions in the real world, e.g. because it was paired with an RL agent, could lead to large-scale irreversible damage, e.g. by accidentally inventing and spreading a lethal pandemic-capable virus while doing medical research. While this might sound like sci-fi, we want to note that a) the difference between harm and help (e.g. poison and cure) could just be one causal mechanism and b) even if the probability is small, the magnitude of the risk justifies the effort to understand the LLMs causal models nonetheless. For a more detailed motivation see [cau, 2022b,a, Everitt et al., 2021].

In this post, we will investigate these causal relationships in multiple very simple settings. One question that we are specifically interested in is whether the LLM bases its answer primarily on the form of the sentence or on its content. For example, does the order in which facts are presented matter for the end result when the content stays the same?

*marius.hobbhahn@gmail.com

ML Safety Workshop, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Table 1: **Cause & effect two sentences:** We present two sentences with a causal relationship and ask the LLM to identify the cause/effect. K-shot setting is omitted for brevity but can be inferred from the 1-shot setting.

Name	Example
Two sentences cause	My car got dirty. I washed the car. Question: Which sentence is the cause of the other? Answer by copying the sentence:
Two sentences effect	My car got dirty. I washed the car. Question: Which sentence is the effect of the other? Answer by copying the sentence:
Two sentences switched	I washed the car. My car got dirty. Question: Which sentence is the cause of the other? Answer by copying the sentence:
Two sentences one-shot	The child hurt their knee. The child started crying. Question: Which sentence is the cause of the other? Answer: The child hurt their knee. My car got dirty. I washed the car. Question: Which sentence is the cause of the other? Answer by copying the sentence:

Table 2: **Further setups:** The "One sentence cause & effect" is taken from BigBench. The other two toy datasets are created to remove the confounder of real-world knowledge and isolate the causality aspect. More details can be found in Appendix A.

Name	Example
One sentence cause & effect	I washed the car because my car got dirty. My car got dirty because I washed the car. Question: Which sentence gets cause and effect right? Answer by copying the sentence
Three balls	The blue ball hit the red ball. The red ball hit the green ball. The green ball fell into the hole. Question: Which ball started the chain? Answer in three words:
Three nonsense-words	The schleep hit the blubb. The blubb hit the baz. The baz fell into the hole. Question: What started the chain? Answer in two words:

We will only investigate the input-output behavior of LLMs in this work but are interested in opening the black box, e.g. by using mechanistic interpretability tools [Elhage et al., 2021] in the future.

2 Setup

We work with four different datasets/setups related to causal understanding. The first two are taken from BigBench [Srivastava et al., 2022], a large benchmark for LLMs maintained by Google. They focus on the plausibility of causal relations in the real world. The third and fourth tasks are toy problems that we created ourselves to isolate causal relationships from world knowledge to prevent confounders (for more details, see Appendix A).

Cause & effect two sentences: Our first BigBench task presents two sentences with a causal relation. The goal is to copy the sentence that reflects the causal relationship correctly (see Table 1). An answer is judged as correct iff the LLM copied the correct sentence.

Cause & effect one sentence: The second BigBench task presents two sentences that have an internal causal relationship and the task is to choose the one sentence that represents the causal info correctly (see 3). An answer is judged as correct iff the LLM copied the correct sentence.

Toy example - 3 colored balls: While the previous cause and effect tasks test for causal understanding in the real world, they also assume some world knowledge, i.e. it is possible that an LLM has a good understanding of causal effects but lacks the world knowledge to put them into place. Therefore, we create simple and isolated examples of causal setups to remove this confounder. We ask questions about the position of the ball in a logical chain and we switch around the sentences such that the

content stays the same but the presentation is different (see Table 3). An answer is judged as correct iff the answer only contains the correct color.

Toy example - 3 nonsense words: The tasks with the 3 colored balls could still require specific knowledge about balls and how they interact. Therefore, we added a variation to the task in which we swapped the colored balls with nonsense words such as baz, fuu, blubb, etc (see Table 5).

Toy example - 5 colored balls: To create a more complicated toy setting, we use 5 colored balls. It is a copy of the 3 colored balls setting except that the chain of balls now contains 5 balls rather than 3. The "switched" condition is now replaced with a "shuffle" condition where the order of colors is randomly chosen rather than switched.

An example of a sentence would be *The blue ball hit the red ball. The red ball hit the green ball. The green ball hit the brown ball. The brown ball hit the purple ball. The purple ball fell into the hole.*

3 Experiments

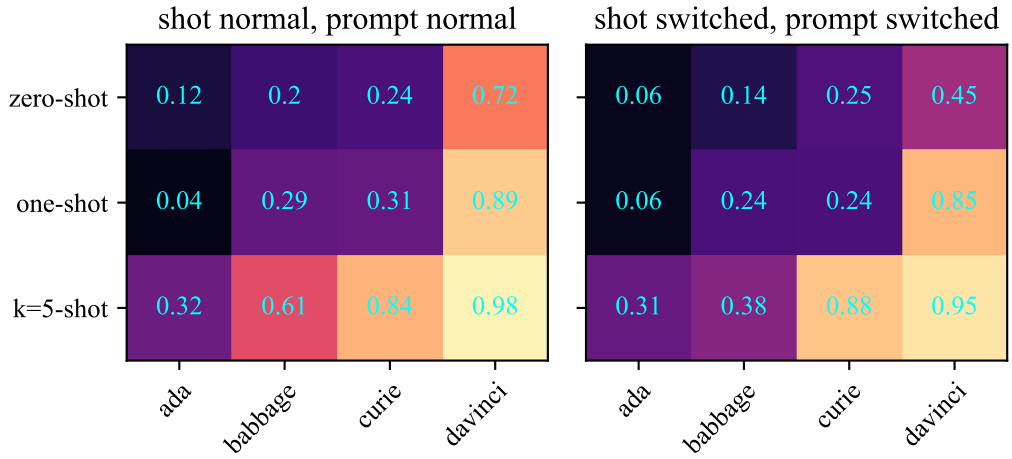

Figure 1: Averaged results across all tasks to isolate the effect of model size and shots. We find that switching the order of the shot and prompt has a large effect on zero-shot performance but has only small effects on one- and k-shot performance.

The main purpose of this work is to identify whether the models understand the causal relationship within the tasks or whether they base their answers primarily on the form of the question. Furthermore, we want to test how the model size and the number of shots (zero, one, k) influence this performance.[2]

3.1 Model size and number of shots

We check the performance on four different versions of GPT-3: Ada, Babbage, Curie and Davinci and three different shot settings: zero-shot, one-shot and k=5-shot. The model sizes are 350M, 1.3B, 6.7B and 175B for the four models respectively. [3]

The detailed results can be seen in Figure 3 and the aggregation in Figure 4. Unsurprisingly, we find that larger models yield better results and that k-shot is better than one-shot and one-shot is better than zero-shot.

3.2 Switch the order in prompts

The LLM might base its answers on the form and not on the content of the prompt, e.g. it could always reply with the first color it identifies rather than answering the prompt. Therefore, we switch the order of the prompts (shots are also switched). The detailed results can be found in Figure 5 and the summary in Figure 1.

[2]link to code will be added upon publication

[3]OpenAI has not made this information public; this information is based on EleutherAI's estimates.

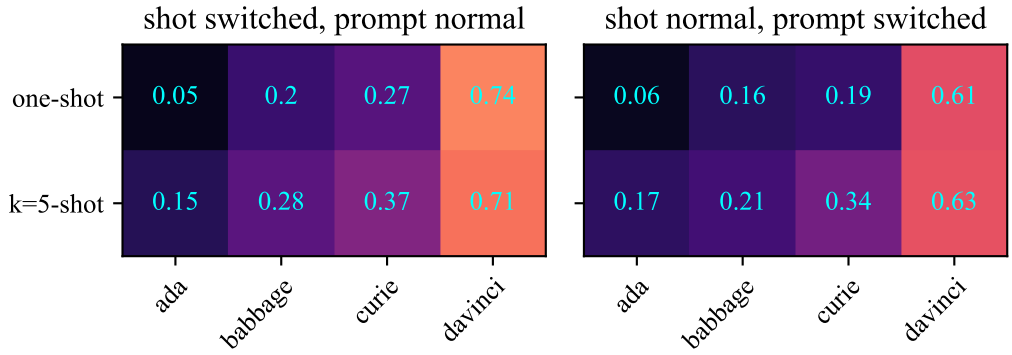

Figure 2: Averaged results across all tasks to isolate the effect of model size and shots. We find that when the order of the shot is different than the order of the prompt, all models perform worse than when they are the same order (see previous figures). One could interpret this as the model being influenced by the form of the sentence and not only its meaning.

The averaged results indicate that switching the order of the shot a) has a large effect on zero-shot performance but b) has only small effects on one- and k-shot performance.

3.3 Switch the order in prompts and shots

The previous results can still be explained if the model focuses more on form than content, e.g. the LLM could still learn to copy the second color it finds rather than the second color in the chain. Therefore, we switch the presentation between shot and prompt, i.e. we present order AB in the one- and k-shots and BA in the prompt. This way, the model has to focus on content rather than form to get the right answer. The detailed results can be found in Figure 6 and the summary in Figure 2.

On average, the scenarios where shot and prompt are in a different order perform worse than when they are in the same order. This would indicate that the model focuses at least to some extent on form rather than content, i.e. it tries to replicate the pattern of the prompt and not its implied causal relationship. This effect seems to be stronger in the cause-and-effect setups than in the toy scenarios.

There is an additional trend that in some scenarios, e.g. first color/word, the k-shot performance is much worse than the one-shot performance. This would indicate that the model is "baited" by the examples to focus on form rather than content, i.e. it is primed on the form by the shots.

When investigating the effect of model size and shots for this setup we find multiple observations. Firstly, all of these results are much worse than in the non-switched setup, i.e. the one- and five-shot performances on Davinci drop by $\sim$0.2 for both. Secondly, for the biggest model, the one-shot results, are better than the k=5-shot results (only for the "shot switched, prompt normal" setting). This would strengthen the hypothesis that the model is "baited" by more shots to focus on form rather than content. An additional experiment with longer chains can be found in Appendix A.

4 Conclusion

We draw three main conclusions from the report. Firstly, larger models yield better results. This has been consistent throughout all experiments and other work and is not surprising. Second, k-shot outperforms one-shot and one-shot outperforms zero-shot in standard conditions, i.e. where shot and prompt have a similar pattern. Thirdly, if the shot and prompt have a different pattern but similar content, this decreases the performance of the model. Furthermore, we find that switching the order of presentation in the prompt decreases the zero-shot performance (see Figure 8).

We expect that LLMs will ultimately be able to solve these tasks based on content and not by pattern matching. For example, our experimental results might look different with bigger models such as PaLM [Chowdhery et al., 2022] (we used GPT-3). However, we think that our results emphasize that LLMs can produce patterns that seem plausible and fit the suggested pattern while being logically incorrect. In these simple toy examples, it is easy to realize that the output is wrong but in more complex scenarios it might not be easy to spot and people could be fooled if they aren't careful. We

therefore want to emphasize that monitoring and understanding causal models in LLMs is one of many steps to increase safety.

References

Causal confusion as an argument against the scaling hypothesis, 2022a. URL https://www.lesswrong.com/posts/FZL4ftXvcuKmmobmj/causal-confusion-as-an-argument-against-the-scaling.

Causality, transformative ai and alignment - part i, 2022b. URL https://www.lesswrong.com/posts/oqzasmQ9Lye45QDMZ/causality-transformative-ai-and-alignment-part-i.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022. URL https://arxiv.org/abs/2204.02311.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Tom Everitt, Ryan Carey, Eric D. Langlois, Pedro A. Ortega, and Shane Legg. Agent incentives: A causal perspective. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(13): 11487–11495, May 2021. ISSN 2159-5399. doi: 10.1609/aaai.v35i13.17368. URL http://dx.doi.org/10.1609/aaai.v35i13.17368.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, and Abubakar Abid. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, 2022. URL https://arxiv.org/abs/2206.04615.

A Appendix

A.1 Prompt engineering

Prompt engineering is still more art than science. We got a better understanding from reading up on the topic, but ultimately trial and error yielded the best results. Specifically, the things we found helpful were:

- No trailing spaces - these just make everything worse for some reason.
- Using a "Question: X? Answer: Y" pattern increased the quality of the output.
- Using "Answer in three words:", "Answer in two words:" or "Answer by copying the sentence" also increased the quality of the output.

We have not run any statistical tests for the above findings, these are our intuitive judgments. So take them with a grain of salt.

Table 3: **Cause & effect one sentence:** We present one sentence with a causal relationship and ask the LLM to identify the cause/effect. K-shot setting is omitted for brevity but can be inferred from the 1-shot setting.

Name	Example
One sentence cause	I washed the car because my car got dirty. My car got dirty because I washed the car. Question: Which sentence gets cause and effect right? Answer by copying the sentence:
One sentence switched	My car got dirty because I washed the car. I washed the car because my car got dirty. Question: Which sentence gets cause and effect right? Answer by copying the sentence:
One sentence one-shot	I washed the car because my car got dirty. My car got dirty because I washed the car. Which sentence gets cause and effect right? Answer by copying the sentence: I washed the car because my car got dirty. Someone called 911 because someone fainted. Someone fainted because someone called 911. Which sentence gets cause and effect right? Answer by copying the sentence:

Table 4: **Three balls toy setup:** We present a chain of balls that hit each other in sequence. In the switched settings, we change the order of the presentation but the logical ordering always stays the same. K-shot setting is omitted for brevity but can be inferred from the 1-shot setting.

Name	Example
Three balls first	The blue ball hit the red ball. The red ball hit the green ball. The green ball fell into the hole. Question: Which ball started the chain? Answer in three words:
Three balls second	The blue ball hit the red ball. The red ball hit the green ball. The green ball fell into the hole. Question: Which ball was second in the chain? Answer in three words:
Three balls final	The blue ball hit the red ball. The red ball hit the green ball. The green ball fell into the hole. Question: Which ball fell into the hole? Answer in three words:
Three balls switched	The red ball hit the green ball. The blue ball hit the red ball. The green ball fell into the hole. Question: Which ball started the chain? Answer in three words:
Three balls one-shot	The blue ball hit the red ball. The red ball hit the green ball. The green ball fell into the hole. Question: Which ball started the chain? Answer in three words: The blue ball. The yellow ball hit the red ball. The red ball hit the green ball. The green ball fell into the hole. Question: Which ball started the chain? Answer in three words:

A.2 Setup - details

More detailed version of the Cause & effect one sentence, three balls toy setup and three nonsense words toy setup can be found in Tables 3, 4 and 5.

A.3 Experiments - details

We have added a random chance level for all tasks in all figures. These are fulfilled if you understand the task but don't understand the causal structure, e.g. in the three colors example you could just choose one of the three named colors. The different shades in the detailed figures indicate zero-, one- and k=5-shots from left to right respectively. The descriptions and findings can be found in the captions of the respective figures.

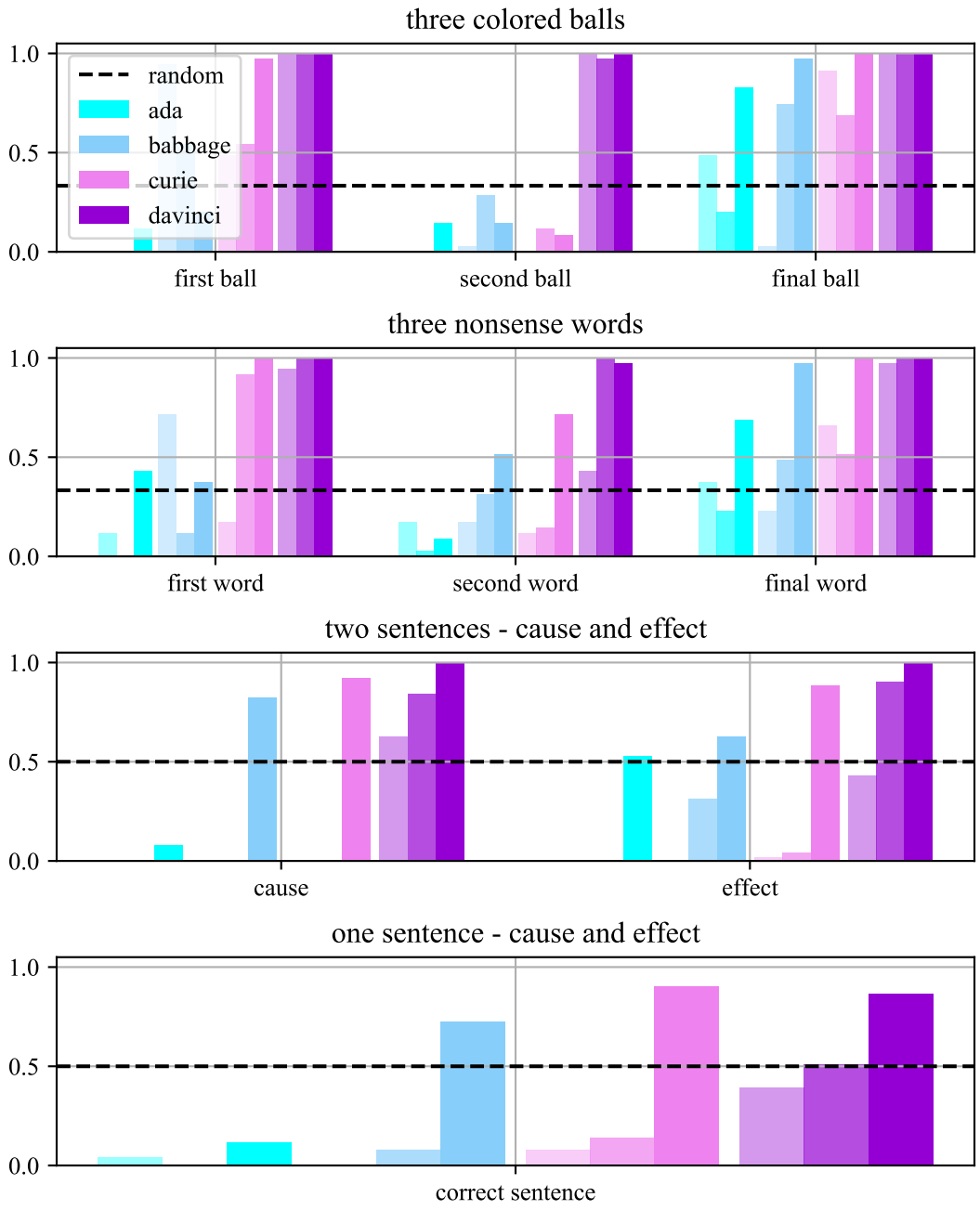

Figure 3: **Comparison of model sizes on four different tasks related to causal reasoning:** The colors indicated different model sizes and the shades indicate zero-shot, one-shot and k=5-shot results from left to right. We find that a) some setups are harder than others, e.g. the toy datasets have better results than the sentences. b) Some tasks are harder than others, e.g. finding the first ball is easier than the second. c) Larger models perform better. d) k-shot performance is better than one-shot and one-shot is better than zero-shot.

Table 5: **Three nonsense words toy setup:** To remove any real-world context from the previous setting, we replace the colored balls of the previous setting with nonsense words. K-shot setting is omitted for brevity but can be inferred from the 1-shot setting.

Name	Example
Three nonsense words first	The schleep hit the blubb. The blubb hit the baz. The baz fell into the hole. Question: What started the chain? Answer in two words:
Three nonsense words second	The schleep hit the blubb. The blubb hit the baz. The baz fell into the hole. Question: What was second in the chain? Answer in two words:
Three nonsense words final	The schleep hit the blubb. The blubb hit the baz. The baz fell into the hole. Question: What fell into the hole? Answer in two words:
Three nonsense words switched	The blubb hit the baz. The schleep hit the blubb. The baz fell into the hole. Question: What started the chain? Answer in two words:
Three nonsense words one-shot	The baz hit the bla. The bla hit the plomp. The plomp fell into the hole. Question: What started the chain? Answer in two words: the baz The baz hit the fuu. The fuu hit the schleep. The schleep fell into the hole. Question: What started the chain? Answer in two words:

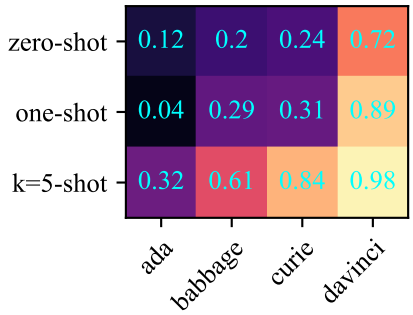

Figure 4: Averaged results across all tasks to isolate the effect of model size and shots. We find that larger models and more shots increase performance in this setting.

A.3.1 Longer chains

In all previous experiments, we have used very simple settings, i.e. we have only switched the order of two sentences. To test if the previous findings generalized to larger settings, we use the 5 colored balls setting. For all experiments, we only used the largest model, i.e. Davinci-text-002 (175B params).

Results on 5 colored balls toy setting can be found in Figure 7. The model seems to always be able to identify the first and last colors. However, the second third and fourth color seem harder to get correctly.

A.3.2 Summary

The main findings of our work can be summarized with Figure 8.

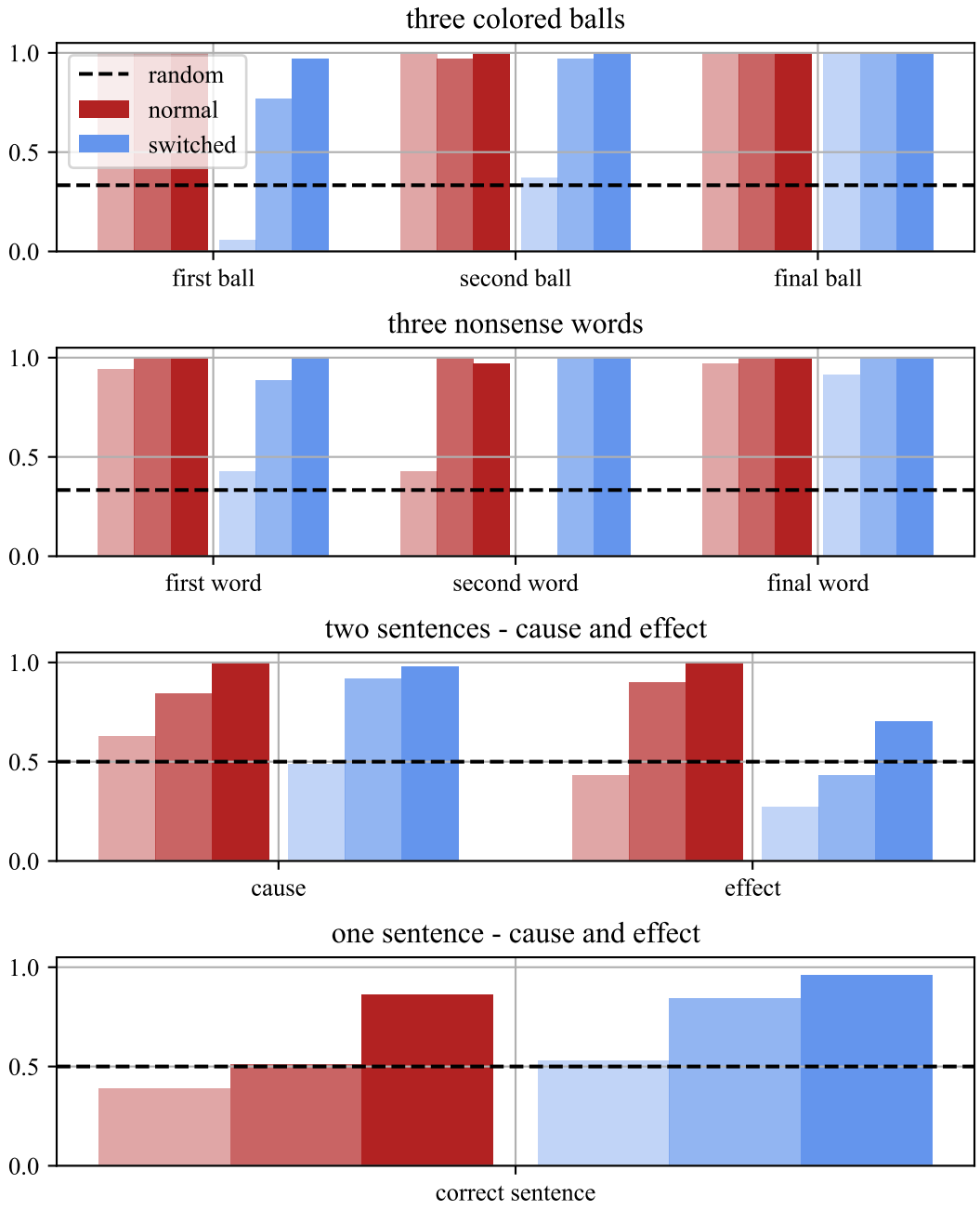

Figure 5: **Effect of switching the order in prompts:** colors indicate different conditions, i.e. normal or switched. The shades indicate zero-shot, one-shot and k=5-shot from left to right. All results are from the largest model (davinci-text-002). We find that switching the order reduces accuracy for the zero-shot settings but not really for the one- and k-shot settings.

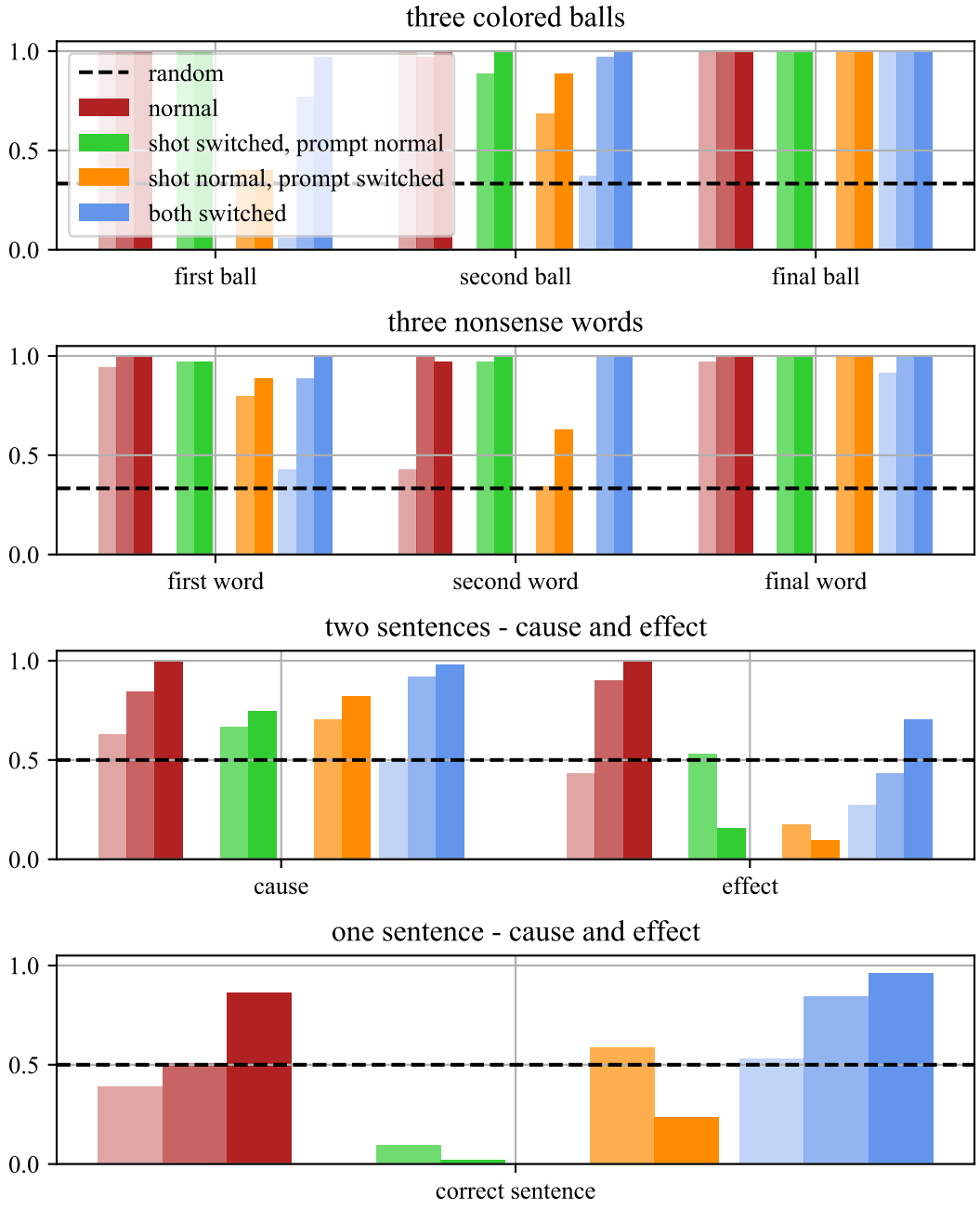

Figure 6: **Effect of switching the order in shots and prompts:** colors indicate different conditions, i.e. combinations of normal and switched shots and prompts. The shades indicate zero-shot, one-shot and k=5-shot from left to right. All results are from the largest model (davinci-text-002). We find that the cross-conditions, i.e. shot switched, prompt normal (or vice versa) perform worse than the unified conditions.

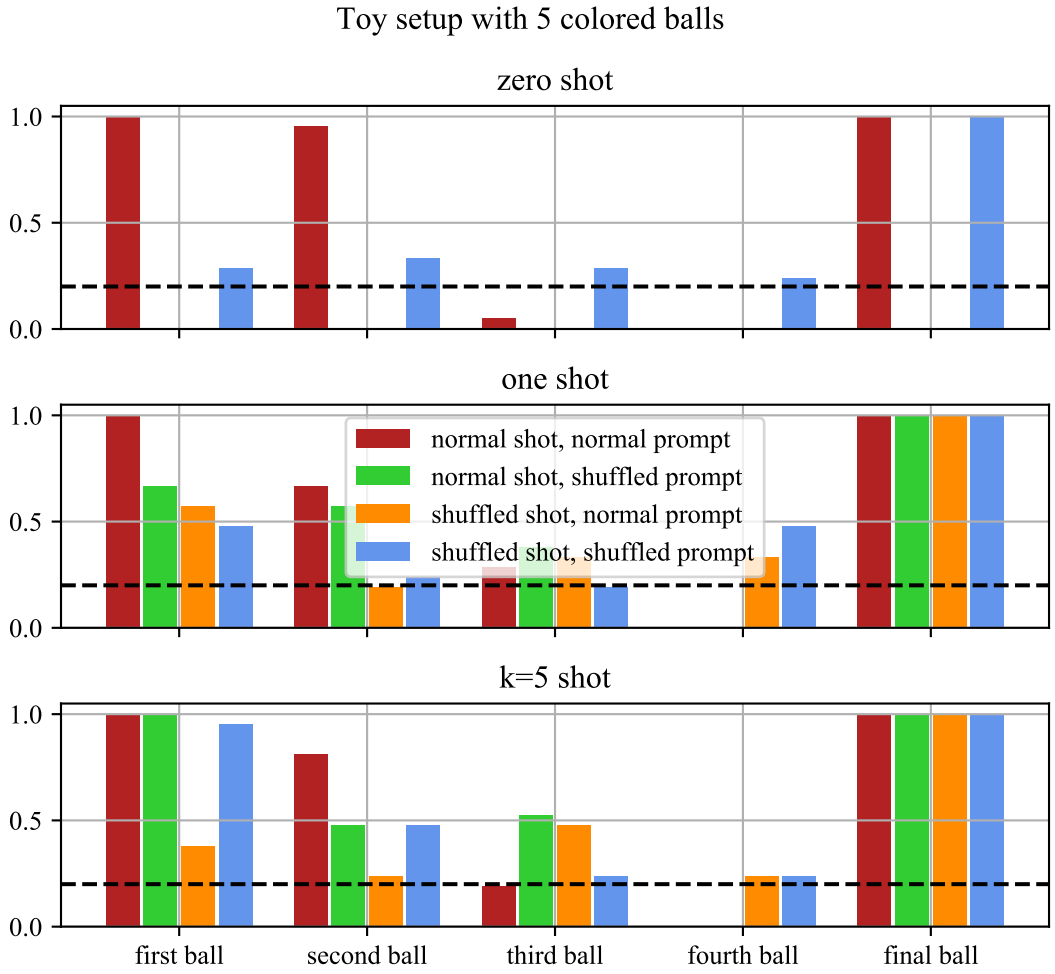

Figure 7: **Results on 5 colored balls toy setting:** The model seems to always be able to identify the first and last colors. However, the second third and fourth color seem harder to get correctly.

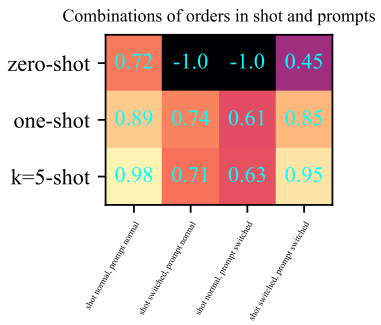

Figure 8: **Comparison of all switched conditions:** The -1 is because you can't switch shot and prompt in zero-shot settings. We find that crossed settings (i.e. columns 2&3) perform worse than ones where shot and prompt follow the same pattern. Furthermore, zero-shot performance is worse in a switched setting.