PROMPT ENGINEERING AND CALIBRATION FOR ZERO-SHOT COMMONSENSE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Prompt engineering and calibration make large language models excel at reasoning tasks, including multiple choice commonsense reasoning. From a practical perspective, we investigate and evaluate these strategies on smaller language models. Through experiments on five commonsense reasoning benchmarks, we find that each strategy favors certain models, but their joint effects are mostly negative.

1 INTRODUCTION

Large Language models (LLMs) have shown impressive performance in many NLP applications (Ouyang et al., 2022; Chung et al., 2022; Wei et al., 2022a), including commonsense reasoning, a key component to AGI (Davis & Marcus, 2015). Recent studies suggest that LLMs are capable of zero-shot and few-shot learning (Brown et al., 2020; Webson & Pavlick, 2022; Chowdhery et al., 2022), and that several strategies can further improve their performance, like prompt engineering and calibration (Kojima et al., 2022; Zhao et al., 2021; Jiang et al., 2021; Kadavath et al., 2022). Despite achieving SOTA performance on many benchmarks, most LLMs are very expensive to use and not released to the public.

Consequently, we study whether prompt engineering and calibration can help smaller language models (those with no more than 3B parameters) in zero-shot multiple choice commonsense reasoning. Since these strategies are likely emergent (Wei et al., 2022b; Chan et al., 2022), we make several modifications, then evaluate them on five commonsense reasoning benchmarks. We find that prompt engineering favors large Flan-T5 models, while calibration works well on GPT-2. Their joint effects are, however, negative in most cases.

2 METHODS

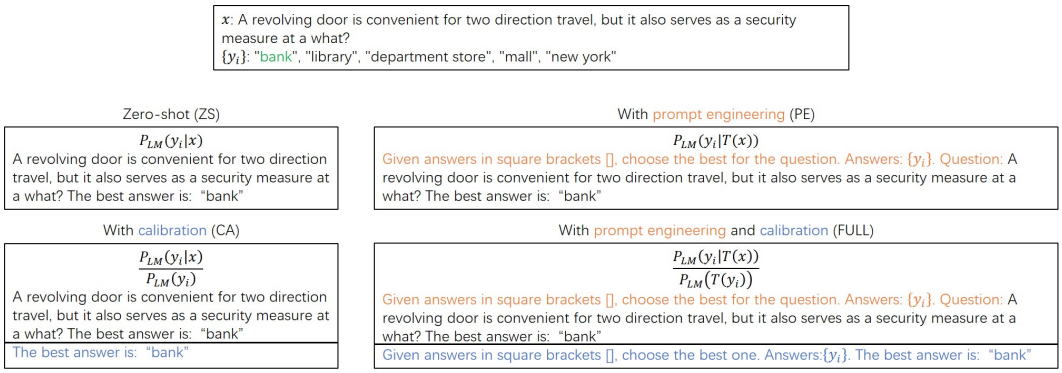

Figure 1: Combinations of data format and option scores for multiple choice commonsense reasoning. Based on the zero-shot method, we add prompt engineering (instruction and multiple choice prompt) and calibration. Unlike previous works, we do not bind options to symbols, like (A).

Background. Multiple choice commonsense reasoning is formalized as follows: Given a question x and several options $y_1, ..., y_n$, select the best option. In the zero-shot setting, a language model

computes a score for each option, which is usually the conditional probability $P_{LM}(y_i|x)$, and selects the one with the highest score, as shown in Figure 1. Recent works suggest that alternatives to the conditional probability can lead to better performance (Holtzman et al., 2021; Niu et al., 2021; Min et al., 2022), but we do not consider these variants for simplicity and fair comparison.

Prompt engineering: multiple choice prompt and instruction. A limit of $P_{LM}(y_i|x)$ is that options are not considered jointly. Recent works suggest that providing all the options in the input, along with instructions about the task, can help LM reason (Robinson & Wingate, 2023; Chung et al., 2022). Inspired by these ideas, we design templates $T()$ that add an instruction and options to a question, as shown in Figure 1. Unlike recent methods that bind each option to a symbol like (A), we use an LM to directly predict answers, because symbol binding is an emergent ability (Robinson & Wingate, 2023).

Calibration. Recent works find that language models prefer certain options even without a question, which suggests they are not well-calibrated (Zhao et al., 2021; Jiang et al., 2021). To overcome this problem, we divide the conditional score of an option by another score computed from a "null" prompt that contains no question, as in $\frac{P_{LM}(y_i|x)}{P_{LM}(y_i)}$. An example is shown in Figure 1.

3 EXPERIMENTS

Datasets. We evaluate prompt engineering and calibration on five multiple choice commonsense benchmarks: (1) CommonsenseQA (CSQA) (Talmor et al., 2019); (2) COPA (Gordon et al., 2012); (3) OpenBookQA (OBQA) (Mihaylov et al., 2018); (4)PIQA (Bisk et al., 2019); (5)Social IQA (SIQA) (Sap et al., 2019); We present their statistics in Appendix B. For all benchmarks, we only use their development sets.

Baselines. We compare four zero-shot methods mentioned in Figure 1: (1) ZS, the standard zero-shot method that computes conditional probability scores of each option; (2) CA, which is ZS with calibration, also known as PMI_{DC} in Holtzman et al. (2021); (3) PE, which is ZS with prompt engineering; (4) FULL, which is ZS with both prompt engineering and calibration.

Setup. As for language models, we use GPT-2 (Radford et al., 2019), T5 (Raffel et al., 2022), and Flan-T5 (Chung et al., 2022), except Flan-T5-XXL, which is too large (11B) to store on our hardware. The evaluation metric is accuracy.

Table 1: Accuracy (%) on Flan-T5

Model	Flan-T5-Small (80M)				Flan-T5-Base (250M)				Flan-T5-Large (780M)				Flan-T5-XL (3B)			
	ZS	CA	PE	FULL	ZS	CA	PE	FULL	ZS	CA	PE	FULL	ZS	CA	PE	FULL
COPA	**59.8**	56.6	52.0	49.6	67.0	**68.2**	60.6	61.4	72.8	71.6	**87.6**	84.0	80.8	78.4	**88.8**	85.6
CSQA	29.2	**37.7**	30.8	28.3	40.9	48.5	**52.5**	51.8	51.6	51.5	62.2	**67.6**	61.8	64.7	70.6	**72.7**
OBQA	14.0	**32.6**	24.8	29.6	20.0	**34.0**	28.6	34.0	24.2	39.4	**53.4**	52.8	30.0	49.6	**61.0**	55.4
PIQA	**62.5**	57.6	54.2	51.1	**65.9**	59.7	58.1	54.0	71.4	65.5	**72.7**	60.6	**75.8**	68.3	68.9	60.4
SIQA	41.7	**42.5**	42.3	42.3	46.4	47.4	**54.7**	53.7	51.4	48.1	**68.6**	66.7	56.1	56.3	**71.6**	58.9

Results. We present results on Flan-T5 in Table 1, and results on GPT-2 and T5 in Appendix C. We find prompt engineering does not work for most models, except the two largest Flan-T5 models, on which it boosts performance by as much as 30 points. This corroborates the effects of instruction-tuning on Flan-T5, and the emergent abilities of larger models (Chung et al., 2022). Apart from that, calibration works well on GPT-2, but inconsistently on other models. This supports findings in Holtzman et al. (2021). Furthermore, the joint effects of both strategies are mostly negative. Overall, our findings suggest careful inspections when using these strategies, as there is no universal configuration that works well on all models.

4 CONCLUSION

We study whether prompt engineering and calibration help smaller language models in multiple choice commonsense reasoning, as they help LLMs. We find that while each strategy favors some language models, their joint effects are mostly negative. Therefore, we suggest careful inspections of these strategies before applying them to smaller language models.

URM STATEMENT

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2023 Tiny Papers Track.

REFERENCES

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. *ArXiv*, abs/1911.11641, 2019.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Stephanie C. Y. Chan, Adam Santoro, Andrew Kyle Lampinen, Jane X. Wang, Aaditya K Singh, Pierre H. Richemond, Jay Mcclelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. *ArXiv*, abs/2205.05055, 2022.

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 M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. 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 García, 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 Díaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *ArXiv*, abs/2204.02311, 2022.

Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc Le, and Jason Wei. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416, 2022.

Ernest Davis and Gary F. Marcus. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, 58:92 – 103, 2015.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pp. 394–398, Montréal, Canada, 7-8 June 2012. Association for Computational Linguistics. URL https://aclanthology.org/S12-1052.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7038–7051, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.564. URL https://aclanthology.org/2021.emnlp-main.564.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, 2021. doi: 10.1162/tacl_a_00407. URL https://aclanthology.org/2021.tacl-1.57.

Saurav Kadavath, Tom Conerly, Amanda Askell, T. J. Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zachary Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, John Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom B. Brown, Jack Clark, Nicholas Joseph, Benjamin Mann, Sam McCandlish, Christopher Olah, and Jared Kaplan. Language models (mostly) know what they know. *ArXiv*, abs/2207.05221, 2022.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=e2TBb5y0yFf.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish 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*, pp. 2381–2391, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1260. URL https://aclanthology.org/D18-1260.

Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Noisy channel language model prompting for few-shot text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5316–5330, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.365. URL https://aclanthology.org/2022.acl-long.365.

Yilin Niu, Fei Huang, Jiaming Liang, Wenkai Chen, Xiaoyan Zhu, and Minlie Huang. A semantic-based method for unsupervised commonsense question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3037–3049, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.237. URL https://aclanthology.org/2021.acl-long.237.

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 E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155, 2022.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1), jun 2022. ISSN 1532-4435.

Joshua Robinson and David Wingate. Leveraging large language models for multiple choice question answering. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=yKbprarjc5B.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 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)*, pp. 4463–4473, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1454. URL https://aclanthology.org/D19-1454.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan 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, Volume 1 (Long and Short Papers)*, pp. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421.

Albert Webson and Ellie 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*, pp. 2300–2344, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.167. URL https://aclanthology.org/2022.naacl-main.167.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022a. URL https://openreview.net/forum?id=gEZrGCozdqR.

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. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022b. ISSN 2835-8856. URL https://openreview.net/forum?id=yzkSU5zdwD. Survey Certification.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12697–12706. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/zhao21c.html.

A Full Prompts for All Benchmarks

In this section, we present prompts (i.e., templates) for each benchmark in Table 2. Specifically, we use one prompt for CSQA and SIQA, and another for COPA, OBQA, and PIQA, because the latter three do not always have a question in a data sample. For simplicity, we still use the term "question" for these three datasets. We also provide the prompts we use for calibration, which is used in FULL.

Table 2: Prompts for each benchmark

Benchmarks	Prompt for the Question	Prompt for Calibration
CSQA, SIQA	Given answers in square brackets [], choose the best for the question. Answers: [*answers*]. Question: [*question*] The best answer is:	Given answers in square brackets [], choose the best one. Answers: [*answers*]. The best answer is:
COPA, OBQA, PIQA	Given answers in square brackets [], choose the one that best completes the sentence. Answers: [*answers*]. Sentence: [*question*] The best answer is:	Given answers in square brackets [], choose the best one. Answers: [*answers*]. The best answer is:

B Dataset Statistics

We present statistics of the five commonsense reasoning (CSR) dataset we use in our experiments in Table 3.

C Results on GPT-2 and T5

We present results on GPT-2 in Table 4, and T5 in Table 5.

Table 3: Statistics of datasets

Dataset Name	Type of CSR	Number of choices	Train	Validation	Test
COPA (Gordon et al., 2012)	Causal	2	N/A	500	500
CSQA (Talmor et al., 2019)	General	5	9741	1221	1140
OBQA (Mihaylov et al., 2018)	Scientific	4	4957	500	500
PIQA (Bisk et al., 2019)	Physical	2	16000	2000	3000
SIQA (Sap et al., 2019)	Social	3	33410	1954	N/A

Table 4: Accuracy (%) on GPT-2

Model	GPT-2-Base (125M)				GPT-2-Medium (350M)				GPT-2-Large (765M)				GPT-2-XL (1.6B)			
	ZS	CA	PE	FULL	ZS	CA	PE	FULL	ZS	CA	PE	FULL	ZS	CA	PE	FULL
COPA	61.0	**62.8**	53.0	54.4	67.0	**70.0**	49.4	54.2	**69.8**	69.4	51.4	57.4	69.0	**71.6**	51.4	53.0
CSQA	25.5	**36.4**	23.8	27.4	30.9	**41.8**	27.4	30.1	33.3	**44.5**	26.9	33.2	38.6	**47.8**	35.1	36.2
OBQA	15.8	**33.4**	25.6	28.0	18.0	**38.6**	26.8	27.4	21.6	**41.4**	25.2	29.4	22.4	**43.2**	25.8	29.4
PIQA	**62.1**	57.1	54.6	52.6	**66.2**	57.5	51.8	52.6	**69.6**	60.7	55.0	54.6	**69.6**	62.2	52.6	53.4
SIQA	35.8	**38.0**	34.3	37.1	36.9	**40.0**	36.0	38.0	36.6	**40.3**	34.0	35.6	39.0	**41.0**	35.2	35.9

Table 5: Accuracy (%) on T5

Model	T5-Small (80M)				T5-Base (250M)				T5-Large (780M)			
	ZS	CA	PE	FULL	ZS	CA	PE	FULL	ZS	CA	PE	FULL
COPA	**55.2**	51.2	51.2	52.2	**59.6**	59.4	51.0	51.8	**65.2**	56.6	53.2	53.8
CSQA	16.6	**22.8**	21.1	21.0	26.1	**30.0**	20.6	22.5	**39.2**	35.4	33.1	35.7
OBQA	14.2	**28.8**	23.8	25.8	15.8	**30.8**	27.8	27.2	19.0	**30.4**	24.8	26.4
PIQA	**56.6**	50.5	51.2	50.8	**61.0**	57.7	51.7	53.0	**66.6**	64.4	52.8	51.7
SIQA	**36.2**	36.1	35.0	34.4	36.2	**37.6**	37.0	33.5	**38.7**	38.1	37.0	34.1

D CODE

Our code is available at `https://anonymous.4open.science/r/Prompt-engineering-and-calibration-0AE0/README.md`