

Multiple Choice Question Answering with Maieutic Prompting

Pasawat Viboonsunti

School of Computing
KAIST

pasawatv@kaist.ac.kr

Rawat Padungkiattiwong

School of Computing
KAIST

rawat.p@kaist.ac.kr

Noppanat Wadlom

School of Computing
KAIST

noppanat_w@kaist.ac.kr

Abstract

We study and introduce an extended application of Maieutic Prompting, a few-shot prompting method designed to address the inherent inconsistency in the reasoning of Large Language Models (LLMs). While existing explanation-based methods, such as Chain-of-Thought prompting, have demonstrated improvements in LLMs’ performance on Multiple-Choice Question Answering (MCQA) benchmarks in unsupervised settings, their reliability is compromised by inconsistent explanations. Our approach extends Maieutic prompting to the MCQA domain by transforming multiple-choice questions into declarative statements and leveraging the original method’s abductive and recursive reasoning within a tree of explanations. To tackle cases where multiple choices are inferred as true, we introduce pairwise abductive reasoning and the concept of pairwise logical integrity. Evaluation on representative MCQA benchmarks, including CommonsenseQA and ARC, demonstrates competitive accuracy compared to state-of-the-art methods like Chain-of-Thought prompting. The results suggest the potential for superior performance due to the robust reasoning steps introduced by our extended Maieutic prompting approach, offering a promising solution to enhance the reliability of LLMs’ reasoning in the MCQA domain.¹

1 Introduction

In the evolving landscape of NLP, MCQA has emerged as a critical domain for evaluating the capabilities of modern language models in the question-answering (QA) task. Recent advances in MCQA have been primarily driven by two approaches: fine-tuning and prompting. The fine-tuning approach in MCQA involves training transformer-based language models, exemplified by the UnifiedQA framework’s use of BART and

T5 models across different QA formats (Khashabi et al., 2020). The prompting approach uses in-context learning with zero-shot or few-shot methods, requiring none or some examples for model inference, with strong performance across various benchmarks (Brown et al., 2020). Additionally, hybrid models like FLAN combine fine-tuning and prompting, showing high effectiveness in MCQA with advancements in task scaling and model size (Wei et al., 2022; Chung et al., 2022).

Besides, the integration of reasoning capabilities in LLMs has become a focal point in advancing their performance, especially in QA tasks. Research has shown that, with causal knowledge, augmenting pre-trained language models enhances their MCQA performance (Dalal et al., 2021). Also, Chain-of-Thought prompting, which offers structured reasoning sequences, enables LLMs to more effectively address complex questions (Wei et al., 2023). Moreover, Maieutic prompting, employing abductive and recursive reasoning, outperforms traditional few-shot methods in binary QA tasks, enhancing LLMs’ capacity for logical and context-rich responses, thereby highlighting the value of advanced reasoning in improving MCQA performance (Jung et al., 2022).

In this work, we aim to apply maieutic prompting to MCQA tasks to explore its effectiveness and potential in dealing with the complexities inherent in multiple-choice scenarios. The Maieutic prompting approach is expected to enhance the logical consistency and contextual relevance of responses from LLMs, contributing significantly to the development of more intelligent and reliable QA systems.

2 Approach

2.1 Maieutic Prompting for MCQA

Our main methodology of Maieutic prompting for MCQA encompasses a two-step process: the gen-

¹Our code is available at <https://github.com/kaitunnz/mcqa-with-maieutic-prompting>

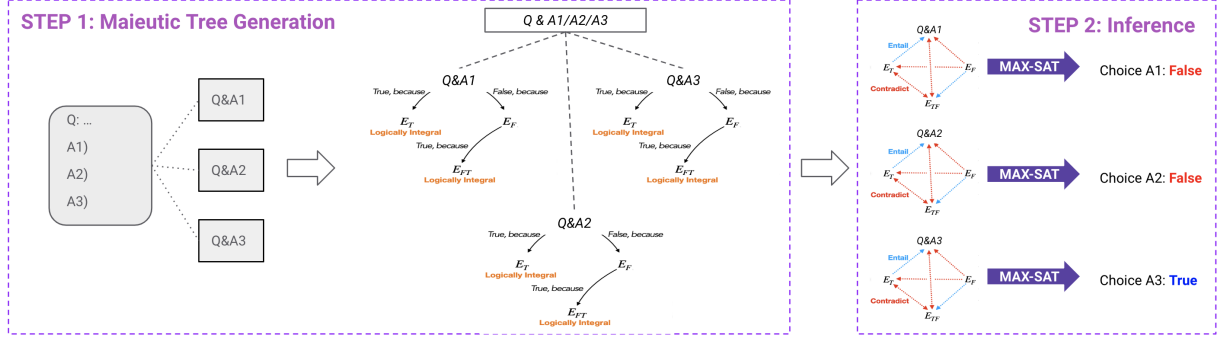


Figure 1: An overview of Maieutic prompting for MCQA. Given a question Q and answer choices A 's, we generate maieutic trees, one for each choice, consisting of abductive and recursive explanations, define the relations between them, and employ MAX-SAT to find the best truth-value assignments to the explanations and Q .

eration of maieutic trees and subsequent inference (Figure 1).

The first step in our approach involves adapting the Maieutic prompting method for MCQA. Initially, we develop new prompts tailored to shift the focus from binary QA to MCQA by generating maieutic trees according to the original algorithm (Jung et al., 2022), with each tree corresponding to a different answer choice for a given question. To effectively reformulate the question statement for Maieutic prompting in the MCQA task, we create two distinct statements for each answer choice. These statements combine the original question with the answer choices, with one statement asserting the answer choice as correct and the other as incorrect. This bifurcation aids in creating a clear, structured approach for the subsequent prompting process.

The second step focuses on inference. After generating the maieutic trees, we process to establish relationships among the different elements using an entailment graph, allowing the logical connections between the question statements and their respective answer choices. Once the entailment graphs are established, we employ a MAX-SAT solver to determine the truth values of the answer choices by analyzing the entailment graphs and deducing which answer choices are most logically consistent and likely to be true. In scenarios where only one choice is deemed as true by the MAX-SAT solver, we consider that choice as the correct answer.

2.2 Pairwise Comparison

In instances where the MAX-SAT solver yields either no true answer or multiple true answers, we implemented a novel *pairwise comparison* technique, illustrated in Figure 2.

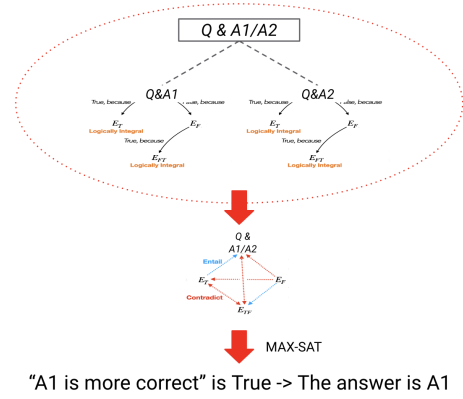


Figure 2: The Pairwise Comparison Technique. Given a question Q and two candidate answer choices $A1$ and $A2$, we assess the pairwise logical integrity to obtain the answer. If not logically integral, we perform pairwise abductive reasoning at depth 1, maieutic tree generation and inference

We initiate a pairwise comparison of the candidate choices, namely $A1$ and $A2$. This process begins by pairing two potential answers and assessing their *pairwise logical integrity* (defined in Appendix A) at the root level, guided by our designed prompts. If the root is pairwise logically integral in this pairwise context, we can deduce the correct choice between the paired options.

In cases where pairwise logical integrity is not established, we employ a technique termed *pairwise abductive reasoning*. This process involves the construction of a single maieutic tree, in which one leaf is the explanation from the combining the question Q and $A1$ statement for the true answer and the other leaf is the explanation from the combining Q and $A2$ statement for the false answer. Then, we perform the traditional maieutic tree generation (including abductive reasoning) from depth 1 onwards, and, after that, conduct the traditional

inference by constructing the entailment graph and applying the MAX-SAT solver as stated in the previous subsection.

3 Data and Experiments

3.1 Replication

Datasets For comparability, similar to the original paper, we evaluate our replication on three binary QA benchmarks for commonsense reasoning including Com2Sense (Singh et al., 2021), CSQA 2.0 (Talmor et al., 2022), and CREAK (Onoe et al., 2021). Only the dev sets of the datasets are used because test labels are not publicly available.

Baselines We compare our replication results to two widely used baseline prompting methods: *Direct prompting*, which prompts the LMs to output only the answers, and *chain-of-thought prompting* (*CoT prompting*). For all prompting methods, including Maieutic prompting, we provide the models with 6 few-shot examples as presented in Appendix C and conduct our own evaluation.

Procedure The replication is divided into two stages: *partial replication* and *full replication*. In partial replication, maieutic trees created by the original paper’s authors for reproduction are used to generate predictions. Following the original paper’s approach, RoBERTa (Zhuang et al., 2021) fine-tuned on MNLI (Williams et al., 2018) is used as a verifier model, and RC2 (Morgado et al., 2014) is used as a MAX-SAT solver.

For full replication, we employ the LMs to generate a maieutic tree for each sample in the datasets, and the inference step is performed on the generated maieutic trees. Unlike the original paper, instead of GPT-3 (*text-davinci-001*), due to the limited availability of API access, we opt for, as the LMs, a more accessible LLM, PaLM 2 (*text-bison-001*) (Anil et al., 2023), and an open-source model, Flan-T5 (Flan-T5-XL) (Chung et al., 2022), which shows performance comparable to GPT-3. The inference step is identical to the partial replication. We also measure the execution time of maieutic prompting to compare with the baseline approaches and fine-tuned Maieutic prompting parameters to be used in the main experiment.

Maieutic Tree Generation with PaLM 2 In Maieutic prompting, the probabilities of output tokens are needed to determine the logical integrity of a proposition in the generation step. However,

Google’s PaLM API, through which we access PaLM 2, does not support querying output token probabilities, so they are approximated with the proportion of times the model says the proposition is true, which we call the model’s *belief* of the proposition, by generating multiple candidate outputs with nucleus sampling. We say that a proposition is logically integral if the absolute difference between the belief of the proposition and that of its negation exceeds a defined threshold.

3.2 Main Experiment

Datasets We evaluate our proposed approach on two MCQA datasets: ARC (Clark et al., 2018), both the Easy Set (ARC-E) and the Challenge Set (ARC-C), and CSQA (Talmor et al., 2019). Only the dev sets are used due to the unavailability of the test labels.

Baselines In addition to Direct prompting and CoT prompting used in the replication, we compare the performance of our approach to Self-Consistency (SC) prompting (Wang et al., 2023), a state-of-the-art prompting method that improves upon CoT prompting by considering multiple reasoning paths. For each baseline prompting method, we provide 7 and 4 few-shot examples on the ARC and CSQA datasets, respectively, as detailed in Appendix C, and conduct our own evaluation.

Configuration Details Due to the poor reasoning ability of Flan-T5-XL observed in the replication results in Section 4.1, we proceed with only PaLM 2 in this experiment. For Maieutic prompting, for each dataset, we manually devised new sets of few-shot prompts for combining a question with an answer choice and a question with the negation of an answer choice, calculating pairwise belief, and performing pairwise abductive reasoning. The prompts can be found in our GitHub repository¹.

We use the Maieutic prompting parameters fine-tuned on the binary QA datasets in the replication process. That is, we set the maximum depth of maieutic trees to 2. For the first depth, we use nucleus sampling with $p = 1.0$ and the temperature of 0.5 to generate 3 E_T s and 3 E_F s. For the second depth, we use nucleus sampling with the same parameters as the first depth to generate 1 E_T and 1 E_F . To calculate the probabilities of output tokens to determine the logical integrity of a proposition, we prompt the model to generate 8 output candidates for the proposition using nucleus sampling

with a temperature of 0.3. The logical integrity threshold is set to 0.2.

4 Results

4.1 Replication Results

Partial Replication Table 1 shows the results of the partial replication. Since the inference step of Maieutic prompting is deterministic, the obtained results are identical to the original paper’s.

Full Replication Table 2 presents the full replication results of Maieutic prompting with PaLM 2 and Flan-T5-XL as the LMs. The Maieutic prompting parameters for both models are fine-tuned on the dev sets, and the optimal parameters for PaLM 2 and Flan-T5-XL are described in Section 3.2 and Appendix B, respectively. It is observed that PaLM 2 demonstrates significantly higher performance than GPT-3 in the original paper, which is expected as it also outperforms GPT-3 on other benchmarks.

In contrast, the results of Flan-T5-XL, also presented in Appendix B, are remarkably less than the original paper’s. To identify the cause, we manually investigated the maieutic trees generated by Flan-T5-XL (omitted for brevity) and discovered that the negations and explanations generated by the model were so incoherent and inconsistent with its answers that they introduced excessive noise which compromised the accuracy of the inference step. The reason is likely that relatively small LMs like Flan-T5-XL are shown to not benefit much from explanation-based prompting (Wei et al., 2023), rendering it unfit for our use case.

Comparison with the Baselines The comparison between the results of Maieutic prompting and the baseline methods using PaLM 2 are shown in Table 3. It can be observed that each prompting method performs differently on different datasets in terms of accuracy with Maieutic prompting achieving the highest accuracy on the Com2Sense dataset. However, for the execution performance, Maieutic prompting falls behind Direct and CoT prompting by 19 times and 14 times on average, respectively.

4.2 Results of the Main Experiment

Table 4 shows the results of Maieutic prompting on the MCQA datasets in comparison with the baseline prompting methods and the SoTA approaches. Overall, we can observe that Direct and SC prompting achieve the best performance on all datasets while Maieutic prompting outperforms

Dataset	Accuracy (%)	
	Original	Replication
Com2Sense	72.5	72.51
CSQA2.0	69.5	69.48
CREAK	85.2	85.27

Table 1: Partial replication results on the three binary QA datasets.

Dataset	Accuracy(%)		
	Orig.	PaLM 2	Flan-T5-XL
Com2Sense	72.5	80.69	61.64
CSQA2.0	69.5	70.97	50.49
CREAK	85.2	89.79	67.76

Table 2: Full replication results with PaLM 2 and Flan-T5-XL on the three binary QA datasets. The results presented are obtained with the prompting parameters fine-tuned on the dev sets.

CoT prompting in terms of accuracy on the ARC-E dataset, demonstrating the viability of Maieutic prompting in the MCQA domain. To understand the performance disadvantage of Maieutic prompting, a closer inspection of the generated maieutic trees is conducted and reveals that individual maieutic trees properly resolve inconsistencies in the model’s reasoning. The actual source of performance loss is identified and explored in Section 5

Furthermore, despite the viability in terms of prediction accuracy, our implementation of Maieutic prompting for MCQA falls behind the other prompting methods in terms of execution performance by a large margin. More precisely, according to Table 4, due to the greater number of prompts required, Maieutic prompting is slower than Direct, SC, and CoT prompting by roughly 160 times, 16 times, and 120 times, respectively, demonstrating the need for massive optimization to be practical.

5 Discussion

5.1 Limitations

Inefficiency of the Initial Maieutic Trees Our approach starts by generating a maieutic tree for each choice of the MCQA questions to infer the truth value of each choice. We then implemented the pairwise comparison technique to deal with cases where none or more than one choice was inferred to be true. We initially believed that pairwise comparison needed to be employed only in a small portion of the questions, but the analysis of the results reveals that, in all the datasets, more than

Dataset	Accuracy (%)			Average samples per minute		
	Direct	CoT	Maieutic	Direct	CoT	Maieutic
Com2Sense	79.03	80.56	80.69	81.6	58.2	4.31
CSQA2.0	73.02	72.98	70.97	77.5	57.0	3.38
CREAK	91.61	91.90	89.79	89.0	62.0	5.55

Table 3: Comparison between the prompting methods with PaLM 2 on the three binary QA datasets. We compare Direct, CoT, and Maieutic prompting for the binary QA task in terms of accuracy and execution performance. The best values on each dataset are highlighted in bold.

Dataset	Accuracy (%)					Average samples per minute			
	SoTA	Direct	SC	CoT	Maieutic	Direct	SC	CoT	Maieutic
CSQA	91.2 ^a	82.88	82.80	81.74	78.30	52.8	4.07	35.0	0.478
ARC-E	95.2 ^b	95.44	95.26	94.74	95.09	56.7	5.67	42.6	0.349
ARC-C	96.3 ^c	88.96	91.30	88.63	85.62	57.0	5.70	42.4	0.247

Table 4: Experimental results of Maieutic prompting and the baseline prompting methods, including Direct, SC, and CoT prompting, with PaLM 2 and the SoTA approaches from ^aDeBERTaV3-large + KEAR (Xu et al., 2022), ^bST-MoE (Zoph et al., 2022), and ^cGPT-4 (OpenAI, 2023) on the MCQA datasets. The best values discounting the SoTA approaches on each dataset are highlighted in bold.

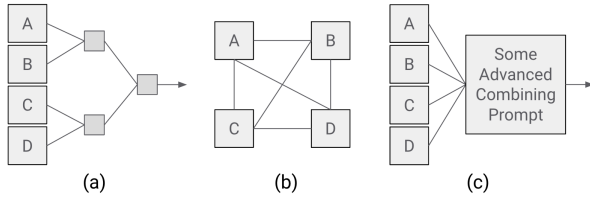


Figure 3: (a) Data path of the current tournament-style pairwise comparison pairing. (b) Data path of a suggested pairing that could reduce inaccuracy. (c) Data path of a hypothetical pairing method that could reduce processing time and inaccuracy

80 percent of the questions were of *edge cases* and required the pairwise comparison technique.

This inefficiency of the initial maieutic tree likely stems from two places. The first is that, in each of the initial maieutic trees, the choices are individually inferred without being contrasted to the other choices. The second is that the prompting parameters were not optimized specifically for the MCQA task. This made the initial maieutic trees infer 50-60 percent of the choices as true.

This unexpected shortcoming of the initial maieutic tree leads to several other limitations of our MCQA method such as requiring excessive amount of trees and having suboptimal accuracy.

Excessive Number of Maieutic Trees Our approach requires an exceedingly large number of maieutic trees for both the initial maieutic trees and the pairwise inference trees compared to just

one for the binary QA task. For instance, in the CSQA dataset whose questions have 5 choices, 5 trees are constructed to infer the truth value of each choice individually, and about 2.5 additional trees are required on average for the pairwise inference to deduce a single correct answer. This means on average, a question in the CSQA dataset will take at least 7.5 times as much time compared to a similarly complex binary QA question.

In Table 4, we see that the number of samples processed per minute is extremely low, about 0.9 percent of the throughput of CoT prompting and 6.9 percent of that of SC prompting. The throughput ratio between Maieutic and CoT prompting is even worse than the 7.5 percent throughput ratio in the binary QA datasets from Table 3. The very low number of samples per minute shown in the results strongly correlates with the excessive number of maieutic trees required.

Suboptimal Accuracy due to Layers of Trees

Our results in the MCQA task exhibit relatively low accuracy compared to the results of Maieutic prompting in the binary QA task in comparison to other methods across datasets. Results analysis have shown that the drop in accuracy stems from the reliance on many layers of maieutic trees.

In the pairwise comparison stage, choices initially inferred to be true are pitted against each other and one is inferred as the more correct answer using a maieutic tree. The choices are paired in tournament-style brackets to infer a single cor-

rect answer. This means the more choices a question has, the more times of pair-up a choice has to go through, and the more pair-up a correct choice has to go through, the more chance it has to be mistakenly eliminated. For instance, if the accuracy of a single maieutic tree is α and there are 4 choices inferred, the chance that the correct answer is not eliminated becomes α^2 in the pairwise comparison stage alone. This also means that if α is low, i.e., the dataset is complicated, the final accuracy will suffer massively.

The fact that more than 80 percent of MCQA questions were the edge case and required pairwise comparison means that the accuracy reduction from the layers of trees affects the accuracy of our method massively.

5.2 Potential Improvements

Optimizing Maieutic Prompting Parameters

As mentioned in Section 3.2, the parameters we used for the MCQA task are optimized on the binary QA datasets. Although we believe that the parameters optimized for binary QA should yield comparable or at the very least decent results, one potential improvement is to further optimize them specifically for this task.

Different Pairwise Comparison Pairing Since our method had accuracy reduced from the layers of trees in the pairwise comparison stage, one improvement is to redesign the pairing method in this stage. Our pairing method currently has the data path shown in Figure 3(a). One potential improvement is to match every single choice together as shown in Figure 3(b), which, although will increase the number of trees and, thus, the processing time, will also improve the accuracy of more complicated questions significantly, as a single inaccurate tree will no longer make the entire inference of the question incorrect.

Combining all the Choices One potential issue of this method is that there is never an instance where all the choices of a question are considered altogether in a single maieutic tree. As seen in Figure 3(c), a prompting method that combines all the choices into one tree will decrease the number of maieutic trees and therefore the processing time required significantly, and the lower number of trees required would mean the accuracy would not suffer from the excessive number of layers of trees.

Preprocessing with Direct Prompting We noticed that, albeit the simplest, the direct prompting method can outperform Maieutic and CoT prompting in multiple datasets. This suggests that some of the questions are straightforward enough that mere Direct prompting can reasonably answer well. Thus, one potential improvement is to preprocess each question with Direct prompting and find if the model is confident in its answer. We can then use the answer from Direct prompting and use Maieutic prompting only if the model is not confident to significantly save processing time. One potential way to probe for the model’s confidence is similar to the way we compute the belief of a proposition as described in Section 3.1.

6 Conclusion

In this project, in addition to replicating the original paper, we explored a way to extend the application of Maieutic prompting from the binary QA task to the MCQA task in order to address the logical inconsistencies in the reasoning of LLMs, a challenge faced by existing methods like CoT prompting. We generate a maieutic Tree to infer the correctness of each choice, eliminating options one by one, and use pairwise abductive maieutic trees to compare the remaining choices and identify the correct answer. Despite promising displays in the reasoning of the maieutic trees, our method demonstrates comparable, rather than superior, performance to state-of-the-art approaches such as CoT and SC prompting on benchmarks like CSQA and ARC.

Acknowledging the potential revealed by our extended Maieutic prompting approach, we recognize several limitations, particularly the substantial computation time required. Another limitation involves the intricate nature of MCQA that induces additive inaccuracy within the many layers of maieutic trees, holding back its decisive edge over existing approaches. Looking forward, possible improvements lie in refining and optimizing elements from framework to prompting parameters and exploring alternative techniques for addressing complexity and inaccuracies. While challenges persist, our study investigated the groundwork for future advancements, exploring ways to push the boundaries of reasoning methodologies for improved QA systems.

References

- Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. [Palm 2 technical report](#).
- 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](#).
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. [Scaling instruction-finetuned language models](#).
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. [Think you have solved question answering? try arc, the ai2 reasoning challenge](#).
- Dhairya Dalal, Mihael Arcan, and Paul Buitelaar. 2021. [Enhancing multiple-choice question answering with causal knowledge](#). In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 70–80, Online. Association for Computational Linguistics.
- Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. [Large language models can self-improve](#).
- Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. 2022. [Maieutic prompting: Logically consistent reasoning with recursive explanations](#).
- Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. [Unifiedqa: Crossing format boundaries with a single qa system](#).
- Antonio Morgado, Carmine Dodaro, and Joao Marques-Silva. 2014. Core-guided maxsat with soft cardinality constraints. In *Principles and Practice of Constraint Programming*, pages 564–573, Cham. Springer International Publishing.
- Yasumasa Onoe, Michael J. Q. Zhang, Eunsol Choi, and Greg Durrett. 2021. [Creak: A dataset for commonsense reasoning over entity knowledge](#).
- OpenAI. 2023. [Gpt-4 technical report](#).
- Shikhar Singh, Nuan Wen, Yu Hou, Pegah Alipoormolabashi, Te-lin Wu, Xuezhe Ma, and Nanyun Peng. 2021. [COM2SENSE: A commonsense reasoning benchmark with complementary sentences](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 883–898, Online. Association for Computational Linguistics.
- 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.
- Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2022. [Commonsenseqa 2.0: Exposing the limits of ai through gamification](#).

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. [Self-consistency improves chain of thought reasoning in language models](#).

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. [Finetuned language models are zero-shot learners](#).

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](#).

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. [A broad-coverage challenge corpus for sentence understanding through inference](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Yichong Xu, Chenguang Zhu, Shuohang Wang, Siqi Sun, Hao Cheng, Xiaodong Liu, Jianfeng Gao, Pengcheng He, Michael Zeng, and Xuedong Huang. 2022. [Human parity on commonsenseqa: Augmenting self-attention with external attention](#).

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. [A robustly optimized BERT pre-training approach with post-training](#). In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. 2022. [St-moe: Designing stable and transferable sparse expert models](#).

A Pairwise Logical Integrity

Let Q represent a question, and A and B represent two distinct candidate answers. We define a relation $R(Q, A, B)$ such that $R(Q, A, B)$ holds if and only if, given the question Q , answer A is more likely to be correct compared to answer B . Conversely, $\neg R(Q, A, B)$ means B is more likely to be correct compared to A . Mathematically, we define this relation as follows:

$$R(Q, A, B) = \begin{cases} T, & \text{if } \operatorname{argmax}_{A_f \in A, B} p_{LM}(A_f | Q, A, B) = A \\ F, & \text{otherwise} \end{cases}$$

A question-answer triplet (Q, A, B) is said to be pairwise logically integral if and only if a language model consistently infers the truth values of $R(Q, A, B)$ and $R(Q, B, A)$ to be logically coherent, i.e., $R(Q, A, B)$ is inferred as True (indicating A is more correct) and $R(Q, B, A)$ is inferred as False (indicating B is less correct), and vice versa.

Formally, we define a boolean function $\text{pairwiseintegral}(Q, A, B)$ as follows:

$$1. R(Q, A, B) = T \text{ and } R(Q, B, A) = F$$

$$2. R(Q, B, A) = T \text{ and } R(Q, A, B) = F$$

$$\text{pairwiseintegral}(Q, A, B) = \mathbb{1}_{\{1 \text{ or } 2 \text{ is satisfied}\}}$$

B Replication Results on Flan-T5-XL

In addition to replicating the original paper’s results on PaLM 2, we performed full replication and evaluated the baseline methods on Flan-T5-XL using the same few-shot examples as detailed in Appendix C to investigate the viability of Flan-T5-XL to be used as the LM in the main experiment so as to dispense with the reliance on external APIs. We also fine-tuned the Maieutic prompting parameters on the dev sets of the three binary QA datasets. Although we decided not to use Flan-T5-XL in the main experiment, the evaluation results are presented in this appendix for curious readers.

Table 5 shows the optimal results obtained with the maximum depth of maieutic trees set to 2. For the first depth, we use nucleus sampling with $p = 1.0$ and the temperature of 0.5 to generate 3 E_T s and 3 E_F s. For the second depth, we use nucleus sampling with $p = 1.0$ and the temperature of 0.3 to generate 1 E_T and 1 E_F . To determine the logical integrity of a proposition, we first prompt the model to answer whether the proposition is true using CoT prompting. The probability of the token "true" in its output is taken as the belief of the proposition. Then, we do the same for the negation of the proposition to get the belief of the negation. The logical integrity of the proposition is determined as described in Section 3.1 with the logical integrity threshold set to 0.45.

Dataset	Accuracy (%)			Average samples per minute		
	Direct	CoT	Maieutic	Direct	CoT	Maieutic
Com2Sense	69.57	70.20	61.64	217.23	48.10	1.76
CSQA2.0	56.20	56.12	50.49	246.20	42.06	1.46
CREAK	82.06	85.85	67.76	328.67	43.30	3.46

Table 5: Comparison between the prompting methods with Flan-T5-XL on the three binary QA datasets. We compare Direct, CoT, and Maieutic prompting for the binary QA task in terms of accuracy and execution performance. The best values on each dataset are highlighted in bold.

C Baseline Prompt Prefixes

Given a statement, determine whether the statement makes sense, and explain the reason.

###

Q: Jane loved to upset her parents with loud noises so she threw a paper plate on the floor?

A: This statement is false, because a paper plate is light and will not make any noise when thrown to the floor.

###

Q: To see stars at night, it is better not to turn on the lights?

A: This statement is true, because Stars are seen more clearly when it's dark.

###

Q: If you want a drink that wakes you up, it's better to look for one with a lot of caffeine rather than protein?

A: This statement is true, because caffeine is a stimulant and will wake you up.

###

Q: It was January in New York so Pat knew that he would see more people at the park rather than in the gym?

A: This statement is false, because it's usually freezing in New York on January, so people would prefer staying indoor rather than going outside.

###

Q: A man who can bench press two hundred pounds can easily lift a small child?

A: This statement is true, because a small child typically weighs way less than 200 pounds.

###

Q: It is a hot day, so Fenton grabbed a big, red popsicle. If Fenton doesn't want to stain the floor, he should stand in the room with the carpeted floor?

A: This statement is false, because if one spills popsicle to the carpet, it will be difficult to clean up because the carpet will absorb it.

###

Table 6: Few-shot examples from [Jung et al. \(2022\)](#) formatted as a prompt prefix for CoT prompting on Com2Sense.

Given a statement, infer whether the statement makes sense or does not makes sense, and explain the reason.

###

Q: To see stars at night, it is better to turn on the lights?

A: This statement is false, because stars are seen more clearly when it's dark.

###

Q: The assassination of a prince led to the first World War?

A: This statement is true, because the assassination of Archduke Franz Ferdinand, heir to the Austro-Hungarian throne in 1914 led to the outbreak of the first World War.

###

Q: If you want a drink that wakes you up, it's better to look for one with a lot of caffeine rather than protein?

A: This statement is true, because caffeine is a stimulant and will wake you up.

###

Q: It was January in New York so Pat knew that he would see more people at the park rather than in the gym?

A: This statement is false, because it's usually freezing in New York on January, so people would prefer staying indoor rather than going outside.

###

Q: It is a hot day, so Fenton grabbed a big, red popsicle. If Fenton doesn't want to stain the floor, he should stand in the room with the carpeted floor?

A: This statement is false, because if one spills popsicle to the carpet, it will be difficult to clean up because the carpet will absorb it.

###

Q: A man who can bench press two hundred pounds can easily lift a small child?

A: This statement is true, because a small child typically weighs much less than 200 pounds.

###

Table 7: Few-shot examples from [Jung et al. \(2022\)](#) formatted as a prompt prefix for CoT prompting on CSQA2.0.

Given a statement, infer whether the statement makes sense or does not make sense, and explain the reason.

###

Q: Christopher Columbus circumnavigated the globe before anyone else?

A: This statement is false, because Ferdinand Magellan's crew were the first to circumnavigate the globe.

###

Q: Stoning is still a current form of capital punishment in certain areas of the world?

A: This statement is true, because Afghanistan and Iran is one of those nations that still practice stoning.

###

Q: Your hair will grow your whole life?

A: This statement is true, because hair never stops growing.

###

Q: A businessperson drives a fire truck to work?

A: This statement is false, because a businessperson does not drive a fire truck, a firefighter drives a fire truck.

###

Q: If a person gets a snakebite, it needs to be treated?

A: This statement is true, because snakebites can be fatal if left untreated and should always be checked.

###

Q: Many people use rail transport in China to travel to Australia?

A: This statement is false, because Australia is a separate continent from Asia, which is not reachable by rail transportation.

###

Table 8: Few-shot examples from Jung et al. (2022) formatted as a prompt prefix for CoT prompting on CREAK.

Given a statement, determine whether the statement makes sense.

###

Q: Jane loved to upset her parents with loud noises so she threw a paper plate on the floor?

A: False

###

Q: To see stars at night, it is better not to turn on the lights?

A: True

###

Q: If you want a drink that wakes you up, it's better to look for one with a lot of caffeine rather than protein?

A: True

###

Q: It was January in New York so Pat knew that he would see more people at the park rather than in the gym?

A: False

###

Q: A man who can bench press two hundred pounds can easily lift a small child?

A: True

###

Q: It is a hot day, so Fenton grabbed a big, red popsicle. If Fenton doesn't want to stain the floor, he should stand in the room with the carpeted floor?

A: False

###

Table 9: Few-shot examples from Jung et al. (2022) formatted as a prompt prefix for Direct prompting on Com2Sense.

Given a statement, infer whether the statement makes sense or does not make sense, and explain the reason.

###

Q: To see stars at night, it is better to turn on the lights?

A: False

###

Q: The assassination of a prince led to the first World War?

A: True

###

Q: If you want a drink that wakes you up, it's better to look for one with a lot of caffeine rather than protein?

A: True

###

Q: It was January in New York so Pat knew that he would see more people at the park rather than in the gym?

A: False

###

Q: It is a hot day, so Fenton grabbed a big, red popsicle. If Fenton doesn't want to stain the floor, he should stand in the room with the carpeted floor?

A: False

###

Q: A man who can bench press two hundred pounds can easily lift a small child?

A: True

###

Table 10: Few-shot examples from [Jung et al. \(2022\)](#) formatted as a prompt prefix for Direct prompting on CSQA2.0.

Given a statement, infer whether the statement makes sense or does not make sense.

###

Q: Christopher Columbus circumnavigated the globe before anyone else?

A: False

###

Q: Stoning is still a current form of capital punishment in certain areas of the world?

A: True

###

Q: Your hair will grow your whole life?

A: True

###

Q: A businessperson drives a fire truck to work?

A: False

###

Q: If a person gets a snakebite, it needs to be treated?

A: True

###

Q: Many people use rail transport in China to travel to Australia?

A: False

###

Table 11: Few-shot examples from [Jung et al. \(2022\)](#) formatted as a prompt prefix for Direct prompting on CREAK.

Q: George wants to warm his hands quickly by rubbing them. Which skin surface will produce the most heat?

Answer Choices: (a) dry palms. (b) wet palms. (c) palms covered with oil. (d) palms covered with lotion.

A: Dry surfaces will more likely cause more friction via rubbing than other smoother surfaces, hence dry palms will produce the most heat. So the answer is (a).

Q: Which factor will most likely cause a person to develop a fever?

Answer Choices: (a) a leg muscle relaxing after exercise. (b) a bacterial population in the bloodstream. (c) several viral particles on the skin. (d) carbohydrates being digested in the stomach.

A: Option (b), bacterial population is the most likely cause for a person developing fever. So the answer is (b).

Q: Which change in the state of water particles causes the particles to become arranged in a fixed position?

Answer Choices: (a) boiling. (b) melting. (c) freezing. (d) evaporating.

A: When water is frozen, the particles are arranged in a fixed position; the particles are still moving for all other options. So the answer is (c).

Q: When a switch is used in an electrical circuit, the switch can

Answer Choices: (a) cause the charge to build. (b) increase and decrease the voltage. (c) cause the current to change direction. (d) stop and start the flow of current.

A: The function of a switch is to start and stop the flow of a current. So the answer is (d).

Table 12: Few-shot examples from [Huang et al. \(2022\)](#) formatted as a prompt prefix for CoT and SC prompting on ARC-E and ARC-C.

Q: George wants to warm his hands quickly by rubbing them. Which skin surface will produce the most heat?

Answer Choices: (a) dry palms. (b) wet palms. (c) palms covered with oil. (d) palms covered with lotion.

A: The answer is (a).

Q: Which factor will most likely cause a person to develop a fever?

Answer Choices: (a) a leg muscle relaxing after exercise. (b) a bacterial population in the bloodstream. (c) several viral particles on the skin. (d) carbohydrates being digested in the stomach.

A: The answer is (b).

Q: Which change in the state of water particles causes the particles to become arranged in a fixed position?

Answer Choices: (a) boiling. (b) melting. (c) freezing. (d) evaporating.

A: The answer is (c).

Q: When a switch is used in an electrical circuit, the switch can

Answer Choices: (a) cause the charge to build. (b) increase and decrease the voltage. (c) cause the current to change direction. (d) stop and start the flow of current.

A: The answer is (d).

Table 13: Few-shot examples from [Huang et al. \(2022\)](#) formatted as a prompt prefix for Direct prompting on ARC-E and ARC-C.

Q: What do people use to absorb extra ink from a fountain pen?

Answer Choices: (a) shirt pocket (b) calligrapher's hand (c) inkwell (d) desk drawer (e) blotter

A: The answer must be an item that can absorb ink. Of the above choices, only blotters are used to absorb ink. So the answer is (e).

Q: What home entertainment equipment requires cable?

Answer Choices: (a) radio shack (b) substation (c) television (d) cabinet

A: The answer must require cable. Of the above choices, only television requires cable. So the answer is (c).

Q: The fox walked from the city into the forest, what was it looking for?

Answer Choices: (a) pretty flowers (b) hen house (c) natural habitat (d) storybook

A: The answer must be something in the forest. Of the above choices, only natural habitat is in the forest. So the answer is (b).

Q: Sammy wanted to go to where the people were. Where might he go?

Answer Choices: (a) populated areas (b) race track (c) desert (d) apartment (e) roadblock

A: The answer must be a place with a lot of people. Of the above choices, only populated areas have a lot of people. So the answer is (a).

Q: Where do you put your grapes just before checking out?

Answer Choices: (a) mouth (b) grocery cart (c) super market (d) fruit basket (e) fruit market

A: The answer should be the place where grocery items are placed before checking out. Of the above choices, grocery cart makes the most sense for holding grocery items. So the answer is (b).

Q: Google Maps and other highway and street GPS services have replaced what?

Answer Choices: (a) united states (b) mexico (c) countryside (d) atlas

A: The answer must be something that used to do what Google Maps and GPS services do, which is to give directions. Of the above choices, only atlases are used to give directions. So the answer is (d).

Q: Before getting a divorce, what did the wife feel who was doing all the work?

Answer Choices: (a) harder (b) anguish (c) bitterness (d) tears (e) sadness

A: The answer should be the feeling of someone getting divorced who was doing all the work. Of the above choices, the closest feeling is bitterness. So the answer is (c).

Table 14: Few-shot examples from [Wei et al. \(2023\)](#) formatted as a prompt prefix for CoT and SC prompting on CSQA.

Q: What do people use to absorb extra ink from a fountain pen?
Answer Choices: (a) shirt pocket (b) calligrapher's hand (c) inkwell (d) desk drawer (e) blotter
A: The answer is (e).
Q: What home entertainment equipment requires cable?
Answer Choices: (a) radio shack (b) substation (c) television (d) cabinet
A: The answer is (c).
Q: The fox walked from the city into the forest, what was it looking for?
Answer Choices: (a) pretty flowers (b) hen house (c) natural habitat (d) storybook
A: The answer is (b).
Q: Sammy wanted to go to where the people were. Where might he go?
Answer Choices: (a) populated areas (b) race track (c) desert (d) apartment (e) roadblock
A: The answer is (a).
Q: Where do you put your grapes just before checking out?
Answer Choices: (a) mouth (b) grocery cart (c) super market (d) fruit basket (e) fruit market
A: The answer is (b).
Q: Google Maps and other highway and street GPS services have replaced what?
Answer Choices: (a) united states (b) mexico (c) countryside (d) atlas
A: The answer is (d).
Q: Before getting a divorce, what did the wife feel who was doing all the work?
Answer Choices: (a) harder (b) anguish (c) bitterness (d) tears (e) sadness
A: The answer is (c).

Table 15: Few-shot examples from [Wei et al. \(2023\)](#) formatted as a prompt prefix for Direct prompting on CSQA.