Leveraging Explanations for Improving In-Context Learning in Commonsense Reasoning Tasks

Yifan Wang
7023035
Language Science and Technology Department
Universität des Saarlandes
yifwang@lst.uni-saarland.de

Abstract

This paper presents a study that investigates the use of explanations in prompts to improve the ability of small pre-trained language models to perform commonsense reasoning tasks. Pre-trained language models have shown remarkable in-context learning ability in natural language processing (NLP) tasks, requiring only a few examples. This characteristic eliminates the need for fine-tuning models, which is computationally intensive and time-consuming. In this study, explanations are introduced to help small pre-trained language models with fewer than or equal to 30B parameters solve a challenging NLP task, namely commonsense reasoning. The study involves comprehensive experiments to examine the effects of various explanatory prompts on models' prediction accuracy and failure rate. Results of the experiments demonstrate that the use of explanations in prompts leads to improved performance in commonsense reasoning tasks, and selected informative examples can lead to further improvements.

1 Introduction

In recent years, the development of Large Language Models (LLMs) has garnered attention in the field of Natural Language Processing (NLP). The scale of language models has increased through training on large self-supervised datasets, resulting in notable advancements in various NLP tasks (Brown et al., 2020; Zhang et al., 2022; Raffel et al., 2020). LLMs with hundreds of billions of parameters demonstrate unique capabilities in understanding and responding to user instructions, which are not observed in smaller models (Wei et al., 2022a). Therefore, In-context learning has emerged as a new paradigm for leveraging LLMs, where models can learn and perform natural language tasks with only a few in-context examples provided in prompts (Min et al., 2021b, 2022; Webson and Pavlick, 2021). This eliminates the need for fine-tuning LLMs for downstream tasks, saving time and computational resources. However, as limited instances are provided in prompts, the question remains as to how to make models learn effectively from such instances.

Explanations, which provide additional information about the reasoning process, have been proposed as a possible approach to improving model performance and generalization ability under in-context learning settings. Recent studies have explored incorporating explanations explicitly by annotating examples with explanations and directly feeding them into the model as prompts, resulting in significant performance improvements in LLMs. Additionally, generated explanations by the model can be used to interpret the model's decisions, providing an advantage in the interpretability of deep neural networks.

Despite significant advancements, LLMs still fall short in tasks involving reasoning, such as arithmetic and commonsense reasoning, compared to human performance (Talmor et al., 2018; Rae et al., 2021). This deficiency in reasoning ability highlights the importance of prompt tuning, particularly in leveraging explanations. In this study, a series of experiments are conducted on a commonsense question answering dataset to examine the effectiveness of incorporating explanations in in-context learning for commonsense reasoning tasks, adapting several settings of explanation utilization.

2 Related Work

2.1 Learning from Explanations

Several techniques have been proposed for incorporating explanations in model training and usage. The motivation for such methods is based on the observation that humans can learn patterns after seeing only a few examples with explanations (Ahn et al., 1992), whereas models typically require much larger training datasets. Earlier works

aimed to extract rules from explanations using logical forms and semantic parsers, which were then fed to the models as additional binary value information (Hancock et al., 2018; Srivastava et al., 2017). Other approaches utilized feature attribution methods to induce models to follow specific human priors, thereby enhancing performance while also reducing biases (Liu and Avci, 2019). With the development of pre-trained language models which are shown to possess basic natural language understanding and processing capabilities, recent studies focus on leveraging natural language explanations directly rather than relying on hand-crafted features (Rajani et al., 2019; Murty et al., 2020; Zelikman et al., 2022). Results demonstrate that language models are capable of generating high-quality explanations, and models perform significantly better when conditioned on these auto-generated explanations.

2.2 Commonsense Reasoning

Although (large) pre-trained language models such as GPT (Brown et al., 2020) and BERT (Devlin et al., 2018) have achieved remarkable success in various NLP domains, even reaching human-level performance, their performance on certain tasks is still unsatisfactory and does not match that of their simpler yet more specialized counterparts (Rae et al., 2021). These tasks include commonsense question answering, symbolic reasoning, and arithmetic questions, which typically require higher levels of reasoning ability. Among these, considerable effort has been directed towards the commonsense reasoning task, which entails understanding information about everyday objects and their interactions (Ye and Durrett, 2022). Several techniques have been proposed to activate the pre-trained knowledge of large language models, including explicitly demonstrating to models the human reasoning process (Wei et al., 2022b).

2.3 In-Context Learning

Increasingly scaled language models have greatly influenced the field of NLP in recent years. As these models may contain hundreds of billions of parameters, fine-tuning them can be difficult or even impractical. Nonetheless, research has demonstrated that large-scale models possess remarkable abilities in comprehending and responding to input prompts. As a result, in-context learning (ICL) has emerged as a novel paradigm in NLP that eliminates the need for fine-tuning. By providing natural language instructions and, optionally, a few examples, models can learn to perform a variety of NLP tasks in a generative fashion, achieving performance on par with state-of-the-art systems. This approach has the potential to save researchers from the burden of requiring vast amounts of training data and lengthy training procedures, and allows non-expert users to customize their own models with minimal learning costs.

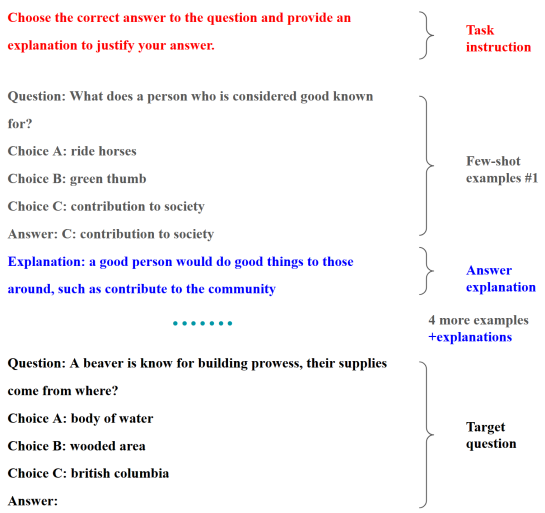

Figure 1: A 5-shot example prompt from CoS-E v1.0 dataset. A default prompt contains task instruction, several examples with explanation and a target question that models are supposed to answer. Performance is evaluated by the frequency of correctly answering questions and failing to choose a valid choice.

Prompt learning has been proposed to fully exploit the knowledge stored in model parameters, in which informative prompts are crafted manually or learned through gradient updates (Liu et al., 2023). Experiments have been conducted to determine the most effective prompts for downstream tasks (Li and Liang, 2021), calibrate model output probabilities (Zhao et al., 2021; Min et al., 2021a), incorporate auxiliary information in prompts (Lampinen et al., 2022), and show models how to reason step-by-step (Wei et al., 2022b). These approaches have led to improved performance, greater stability, and enhanced generalization abilities, even in previously challenging tasks. This indicates the importance of prompt tuning when adopting in-context learning.

3 Methods

In this study, I adopt several different settings of incorporating explanations into prompt, and compare

their performance with the setting without explanations. An instance in our example pool consists of four parts, namely question, choices, answer and explanation. Each part is split by "
n" and preceded by a word indicating its role ("Question:", "Choice:", "Answer:" and "Explanation:") in the instance. Several examples in this form are sampled, and they follow a task instruction to construct a prompt, which is followed by a test instance where no explanation or answer is provided. An example of such prompt is shown in Figure 1. Models are expected to choose the correct option and generate a reasonable explanation to justify their decision. Prediction accuracy of answers is used to evaluate model performance and compared against settings where no explanation or control explanations are used in the prompt. To accurately study the effect of explanations, I also construct prompts with crafted control explanations to eliminate the interference of confounding factors like prompt length and content. Following (Lampinen et al., 2022), two types of control explanations are introduced:

- **Scrambled explanations:** To ensure that model does not only rely on word level information, the word order of explanations is first shuffled before they construct prompts.

- **Other item explanations:** It is also important to determine whether the model benefits from the direct relationship between the question and its explanation rather than any random explanation. In the few-shot cases, explanations are permuted such that no question is matched with its original explanation.

Aside from explanations, effects of the following factors are also evaluated in this experiment and their interactions with explanations are discussed in later sections.

- **Model size**: Previous research has suggested that only large-scale language models are capable of comprehending user intentions and generating corresponding responses, whereas smaller models often struggle in this regard (Rae et al., 2021). In light of this, I conduct experiments on models of varying sizes to investigate the extent to which explanations can mitigate this issue and provide benefits for language models across a range of scales.

- **Number of examples in each prompt:** While it is commonly assumed that few-shot learning leads to better performance than zero-shot learning, in this study I construct a set of prompts with 0-5 examples, aiming to analyze whether an improvement of performance can be observed with more examples added in prompts.

- **Order of explanation and answer:** While large models can arguably better predict answers by conditioning on explanations generated by themselves (Wei et al., 2022b), smaller models might be prone to accumulating errors when generating explanations, leading to poorer overall performance. To investigate this, I conduct experiments in two distinct settings: "answer first," in which models generate an answer and then provide explanations to justify it, and "explanation first," in which models generate explanations prior to making a prediction in a step-wise fashion. The order of answer and explanations in the prompt examples is adjusted accordingly.

- **Instructions:** To evaluate a model's capacity to comprehend intentions, I conduct a comparison between 0- and 5-shot in-context learning performance with and without instructions. Prior to the evaluation, the instruction is manually tuned on a held-out validation set.

- **Sampling/selecting examples:** Examples can exhibit varying degrees of usefulness in terms of their ability to assist a model in generating accurate predictions. In addition to randomly sampling examples from the training set, a greedy search procedure is employed to identify the best 1-, 2-, 3-, 4-, and 5-shot examples, each of which is created by sequentially adding the most effective example to the set of previous best examples. The selection of instances is based on their performance on the validation set.

4 Experiments

4.1 Dataset

This work is mainly devoted to studying the effects of explanations in common sense reasoning tasks, therefore I choose CoS-E v1.0 dataset (Rajani et al., 2019), a commonsense question answering dataset annotated with explanations. The CoS-E dataset

contains a training set of 7610 instances and a validation set of 950 instances, where each instance is a commonsense question with three choices. The annotations for answers and explanations are provided by human annotators. In order to ensure fairness and efficiency, a subset of 100 instances is randomly selected from the training set, from which prompts are sampled for the study. Additionally, another 100 instances from the training set are utilized for validation purposes to optimize hyperparameters. Finally, the overall performance is evaluated and reported on the complete validation set.

4.2 LMs

In order to investigate the impact of model scale on the in-context learning performance and ability to leverage explanations, a series of experiments were conducted using transformer-decoder (Vaswani et al., 2017) only models of varying sizes. Specifically, Specifically, a set of Facebook Open Pre-trained Transformer (OPT) models (Zhang et al., 2022) are utilized, with the number of parameters ranging from 125M to 30B, where the latter size is chosen due to GPU memory limitations. Further details regarding the model architecture and pre-training process can be found in the original publication. During inference, a temperature of 0.9 is set to introduce a degree of randomness in the generated output, and a maximum of 60 tokens are allowed to be generated. A single run of the largest 30B model takes approximately 4 hours using 6 V100 GPUs, each having a memory of 32 GB.

4.3 Evaluation Metrics

As it is challenging to extract explanations directly from model outputs, the performance evaluation in this study is solely based on the model's predictions, instead of the automatically generated explanations. The accuracy in answering the target questions is employed as the evaluation metric to assess the model's proficiency in common sense reasoning. To minimize the possibility of false positive and false negatives, the answers are formatted as "Letter: Answer" (e.g., B: Valley) instead of using only a letter (e.g., B) in this study. In zero-shot settings where the model has not encountered such a format, only the first generated token is used to match the letter corresponding to the correct answer. Additionally, it is observed that language models sometimes fail to generate valid predictions (e.g., failing to answer the question or predicting a non-existing choice). This phenomenon is more common in smaller models, indicating their poorer ability to comprehend instructions and questions. To quantify this occurrence, failure rate is adopted as an additional metric, measuring the frequency of the model failing to generate a prediction belonging to the question choices. The results are averaged over two runs with different random seeds.

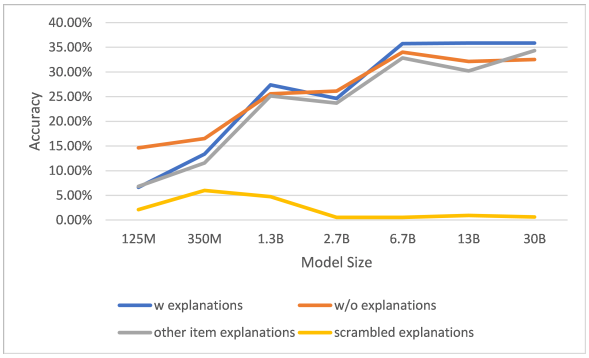

Figure 2: Prediction accuracy of models of various scale with different type of explanations. Larger models are capable of effectively leveraging explanation information, yielding more accurate predictions than counterparts with no or control explanations.

5 Results

5.1 Quantitative Analysis

The primary results of the experiment are presented in Figure 2 and Figure 3. In these 5-shot scenarios, the default configuration is utilized, where no greedy search of examples is conducted, instructions are provided, and answers are given before explanations. The prediction accuracy results in Figure 2 demonstrate that the effects of introducing explanations vary across models of different sizes. While smaller models (125m, 350m) show decreased performance when additionally conditioned on explanations, larger models benefit from additional information from explanations and exhibit higher accuracy. All models achieve better performance with true explanations than with control explanations, suggesting that the improvement indeed comes from the explanations. However, it is important to note that CoS-E contains 3-way multiple choice questions, indicating that an accuracy of roughly 33% can be achieved through random guessing. Even under the best conditions, the largest model in the experiment does not significantly surpass this benchmark, indicating that 30B models still struggle with pure in-context learning.

On the other hand, Figure 3 shows even though

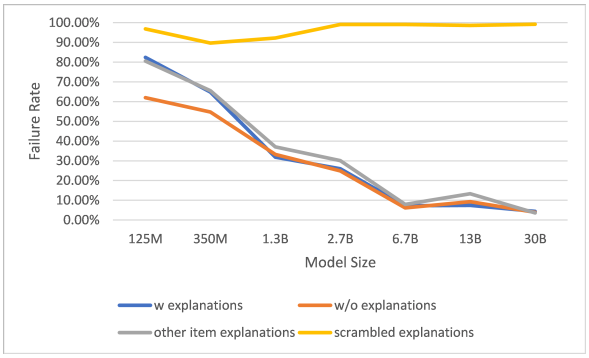

Figure 3: Failure rate of models of various scale with different type of explanations. Despite greater input length and additional information in the prompt, larger models show no significantly worse performance in terms of making valid predictions.

more data is fed into the model before the target question, it does not hurt model's performance of making valid predictions when the model is large (larger than 350m). Conversely, smaller models tends to generate irrelevant responses to the question when explanations are included, potentially because they struggle to establish connections between explanations and the questions. As observed for the accuracy metric, the use of true explanations tends to lead to better failure rates than the use of control explanations.

When comparing the effects of two types of control explanations, it is evident that other item explanations have only a slightly worse effect on performance compared to true explanations, while scrambled explanations have a wholly negative impact on models. The latter hinders the models' reasoning ability, leading them to generate irrelevant responses in almost all cases. One possible explanation for this is that the models rely more on syntactic correctness than semantic relevance and do not recognize scrambled explanations as valid placeholders. While there is an overall trend of larger models having better prediction accuracy and lower failure rate, these metrics do not vary continuously with model size see Table 1). Instead, there is a significant improvement of about 10% in accuracy between 350M and 1.3B models, followed by relatively stable performance for 1.3B and 2.7B models. Similarly, 6.7B, 13B, and 30B models (with 3.48 times more parameters) demonstrate roughly equal prediction accuracy, while a substantial leap occurs from 2.7B to 6.7B models (1.48 times more parameters). This phenomenon may provide further evidence of the emergent ability described in (Wei et al., 2022a; Ganguli et al., 2022).

	Parameter increase ratio	**Accuracy increase ratio**
350M	180.00%	101.66%
1.3B	271.43%	104.71%
2.7B	107.69%	-10.01%
6.7B	148.15%	45.11%
13B	94.03%	0.28%
30B	130.77%	0.00%

Table 1: Parameter and accuracy increase ratio relative to previous models. It can be observed that performance increase is not always to be expected when model is scaled up.

5.2 Qualitative Analysis

A qualitative analysis of model performance is conducted by examining examples where the model generates undesired responses (Table 2). Each example represents a distinct type of error. The first example highlights the model's weakness in reasoning abilities, as the generated explanation is factually correct but irrelevant to the question. In the second example, the model generates a useful explanation containing the correct answer, yet it still predicts the wrong option. This issue is present in both explanation first and answer first settings. The third example shows the model's vulnerability to factual errors in explanations, leading to incorrect predictions. In contrast to example two, the fourth example produces an accurate prediction but with an irrelevant explanation. The last example illustrates situations where no helpful information is contained in the explanation. These examples reveal that the model struggles to generate relevant explanations, and predictions are not always consistent with the explanations generated. The impact of both factors will be discussed in section 6.5.

6 Ablation Studies

6.1 Number of Examples

In addition to analyzing the impact of explanations on model performance, this study also explores the effect of varying the number of examples in the prompt. The results are presented in Figure 4 and Figure 5. Notably, the experiment is limited to models smaller than 6.7B due to computational constraints. Surprisingly, the findings reveal that

Question: A child wants to play, what would they likely want? Choices: **play tag**, breath, fall down Explanation: breathing is what kids do.
Question: What was the man encouraged to do after he expressed his anger violently? Choices: illustrate point, punch, **cool off** Explanation: the man was encouraged to cool off because he expressed his anger violently
Question: What region of a west coast U.S. city would you find a Japanese restaurant? Choices: california, tokio, **downtown** Explanation: tokio is a city in japan. japan is located in west coast of united states
Question: What do you do when you're going to market? Choices: see other people, traveling, **buy food** Explanation: going to market requires food
Question: What do you go to see for live entertainment? Choices: movie, **show**, theatre Explanation: movie

Table 2: Examples of undesired responses. Underlined are model predictions and bold indicates ground truth.

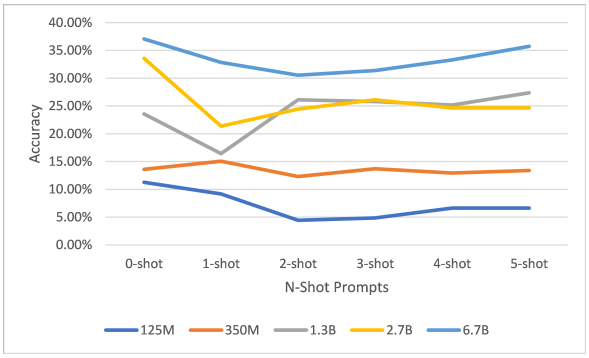

Figure 4: Prediction accuracy of models with n-shot examples as input. 5-shot prompts don't always generate best responses, and 0-shot prompts have the highest accuracy in most cases.

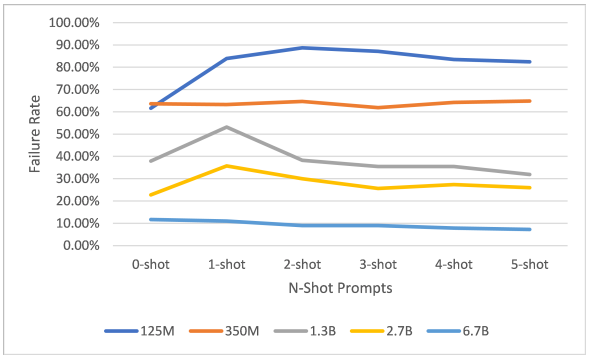

Figure 5: Failure rate of models with n-shot examples as input. Similar patterns to accuracy curve can be seen here. Note that since a different method is adopted to detect failure in 0-shot setting, the failure rate of 0-shot setting can be underestimated.

increasing the number of examples in the prompt does not necessarily improve the model's prediction accuracy or failure rate. In fact, the 0-shot prompt often yields higher or comparable accuracy compared to the 5-shot prompt, indicating that models may not benefit from additional examples. It is hypothesized that this could be due to the models' weak ability to learn patterns from the given examples, resulting in confusion rather than clarity. However, further research is required to confirm or reject this hypothesis, particularly on larger models.

6.2 Effect of Using Instructions

The results regarding using instructions are presented in Figure 6 and Figure 7. The impact of instructions varies according to the model size and number of examples. In zero-shot prompts where no example is presented to the models, the inclusion of instructions leads to better accuracy and lower failure rate for all models. The same effects can also be noticed in 5-shot prompts, but only in larger models (larger than 1.3B). On the other hand, smaller models tend to generate worse responses when instructions are incorporated. In general, in zero-shot learning, the use of instructions always provides better outcomes, and only larger models can benefit from instructions when several examples are already present in the prompt, probably due to their better ability to remember longer context and comprehend instructions.

6.3 Answer First vs. Explanation First

The differences between answer first and explanation first settings are two-fold: Firstly, the order in which answers and explanations are given in the prompt can different. Besides, under explanation first settings, models predict an answer conditioning on the explanation generated by themselves. I study the performance under both settings and present the results in Figure 8 and Figure 9. In general, answer first settings constantly outperform explanation first counterparts in both accuracy and failure rate, with the only exception observed in 125M model. This observation can be probably

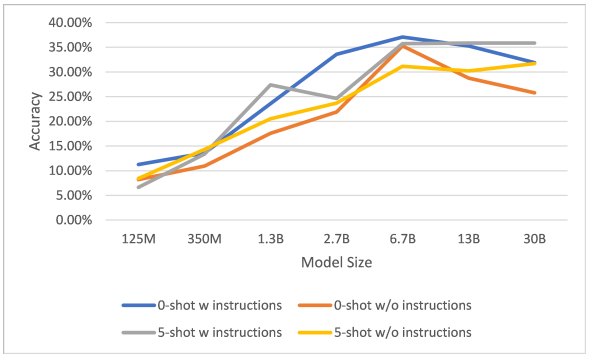

Figure 6: Prediction accuracy of models with and without instructions. Including instructions leads to constant improvement over without explanation setting except for smaller models with 5-shot prompts.

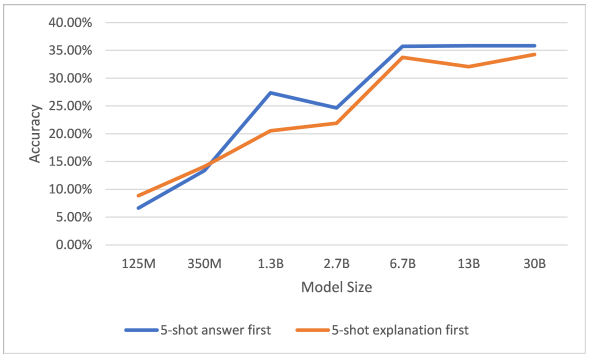

Figure 8: Prediction accuracy of answer first and explanation first models. First answering questions then justifying it usually outperforms explanation first counterparts.

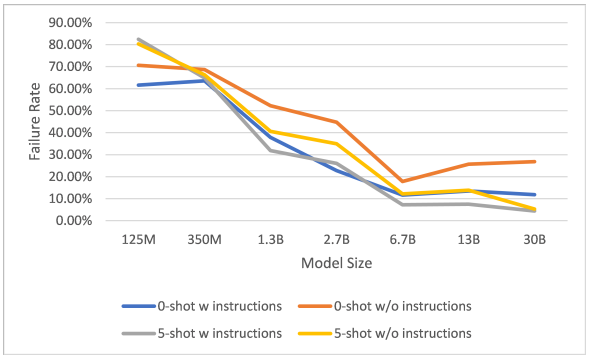

Figure 7: Failure rate of models with and without instructions. Curves show similar trends as in the figure of accuracy.

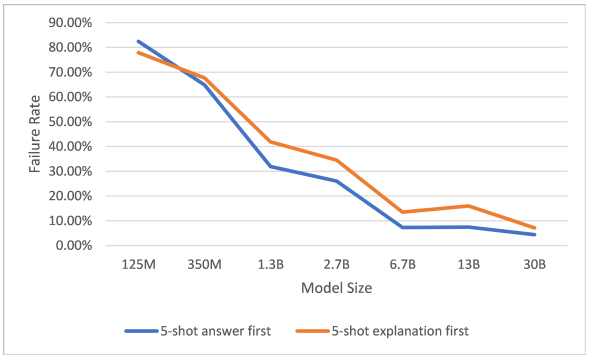

Figure 9: Failure rate of answer first and explanation first models. Making predictions directly after prompts yields lower failure rate, indicating that conditioning on generated explanations induces more noise.

attributed to models' inability of generating reasonable explanations. Conditioning on explanations with noise accumulates error and leads to lower accuracy and higher probability of making invalid predictions. The observed pattern is consistent with (Lampinen et al., 2022), where authors argue that post-answer explanations bring benefits in performance and inference time cost. However, (Wei et al., 2022b) shows different results, where answer first performance does not match chain-of-thoughts first setting and is only about the same as baselines. Some other work (Zelikman et al., 2022; Nye et al., 2021)also adopt explanation first strategies, while more detailed comparative experiments are required to make a final conclusion.

6.4 Greedily Selected Examples

During the experiments, I sample 100 examples to form a candidate pool, and another 100 examples reserved for validation. To determine the optimal 5-shot prompts, I employ a greedy search approach. Specifically, I initially search for the best 1-shot prompt by iterating over all candidates and selecting the prompt with the highest accuracy on the validation set. I then add the remaining examples to the chosen 1-shot prompt and select the 2-shot prompt that yields the best performance. Following this process, I am able to approximate the best 5-shot prompts on the validation set. Since this is approximately equal to making 5 $\times$ candidate pool size $\times$ validation set size inferences, which is computationally very intensive, I only run the search process on 125M, 350M, 1.3B and 2.7B models. For models larger than 2.7B, I use the set of examples obtained from 2.7B models. Results are given in Figure 10 and Figure 11. It shows that greedily searched examples help almost all models make more accurate predictions, even those that are not obtained from the model itself (6.7B, 13B). However, the impact on failure rate is not as pronounced, with the two curves largely overlapping across all models. I thus conclude that greedily searching examples enhances model prediction accuracy, but

Figure 10: Prediction accuracy of models with and without greedy search examples. Greedy search usually returns examples that help improve model performance.

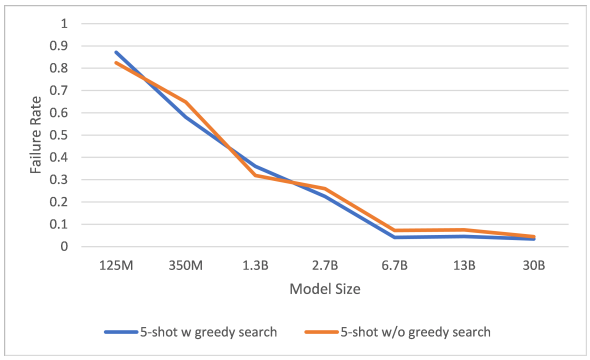

Figure 11: Failure rate of models with and without greedy search examples. Selected examples does not help lower the failure rate of models.

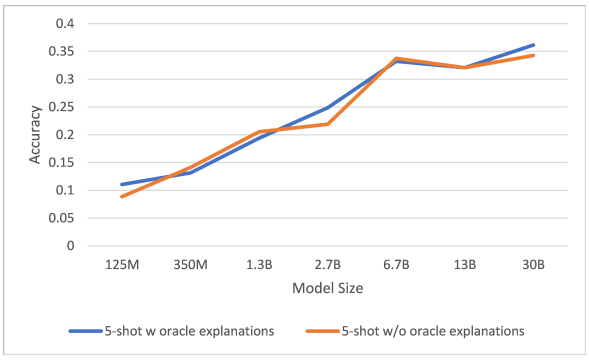

Figure 12: Prediction accuracy of models with and without oracle explanations. No significant difference is observed between two curves.

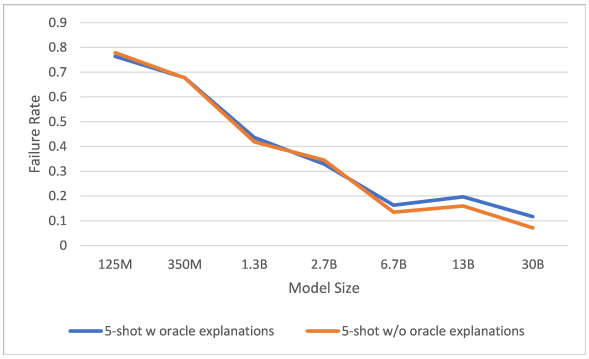

Figure 13: Failure rate of models with and without oracle explanations. No significant difference is observed between two curves.

cannot mitigate invalid predictions. Notably, it is observed that models of varying sizes selects entirely different sets of examples as the best prompts, with few common examples. No clear features have been found between these selected examples through manual inspection, suggesting that future research could explore the type of examples and explanations that benefit few-shot in-context learning the most.

6.5 Explanation Quality

The current experiment explores various methods to improve the performance of a model in predicting correct answers, however, the achieved accuracy is still unsatisfactory and is only equivalent to random guessing. The low accuracy could be attributed to two factors: the model's inability to generate informative explanations or its inability to correctly utilize generated explanations. To determine which factor is responsible, the model is fed with oracle explanations of the target question. If the performance improves, then the bottleneck in the process is the generation of useful explanations.

On the other hand, if no significant difference is observed in the model's performance, then the low accuracy can be attributed to its inability to utilize correct explanations. The results of this experiment are presented in Figure 12 and Figure 13, and it can be concluded that the key factor responsible for the model's poor performance is its inability to fully utilize generated explanations. It is important to note that the oracle explanation models are compared against explanation-first models, as they make predictions by conditioning on explanations, just like the explanation-first models.

Nonetheless, it should be noted that the quality of explanations can still have an impact on prediction accuracy. To assess this impact quantitatively, a correlation test was conducted between the quality of explanations (measured using BERTscore with respect to oracle explanations) and the accuracy of the prediction. The results, presented in Figure 14, demonstrate a significant correlation ($p = 5.95 \times 10^{-8}$) with a Pearson correlation coefficient of 0.744.

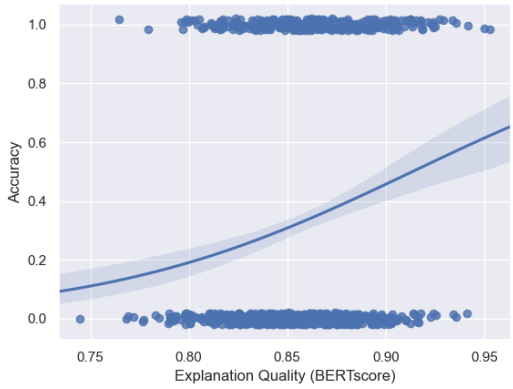

Figure 14: Effect of explanation quality on prediction accuracy. Data points are fitted using a logistic regression curve.

7 Limitations

Following factors limit the reliability of results in this experiment, thus it is questionable whether they can apply to other datasets and models.

- **Too few experiments:** The present study's results are based on the average of two experimental runs conducted with different random seeds, which is insufficient to draw a reliable conclusion. Furthermore, several crucial statistical parameters, such as variance and confidence interval, are absent from the analysis. To confirm the existence of the factors' effects, it is necessary to perform a significance test.

- **Too small models:** Previous research has indicated that models typically exhibit a significant improvement in their ability to perform commonsense reasoning only when they have more than 60 billion parameters (Wei et al., 2022b). However, the largest model used in this study is smaller than this threshold. For comparison, GPT-3 Davinci-003 model can achieve an accuracy of 84.72% without any greedy search. Therefore, the findings of this study may not be generalizable to large language models.

- **No interaction among factors:** To more comprehensively study the effects of each factor, it would be more rigorous to conduct a mixed linear effect analysis that allows for interaction among the factors. Currently, in this experiment, factors are only analyzed in isolation for the sake of simplicity.

8 Conclusion

This study provides empirical evidence demonstrating that incorporating explanations into prompts can augment the in-context learning capability of language models in performing commonsense reasoning tasks. Moreover, the benefits of this approach can be further amplified by employing greedily searched examples. Nonetheless, the advantages of explanations are not noticeable in smaller models. It should be noted even the largest model utilized in this study fails to surpass random guess, even with the aid of explanations. Ablation analyses reveal that this is primarily attributed to the models' incapacity to appropriately leverage the explanations to facilitate reasoning.

References

Woo-kyoung Ahn, William F Brewer, and Raymond J Mooney. 1992. Schema acquisition from a single example. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(2):391.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, et al. 2022. Predictability and surprise in large generative models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1747–1764.

Braden Hancock, Martin Bringmann, Paroma Varma, Percy Liang, Stephanie Wang, and Christopher Ré. 2018. Training classifiers with natural language explanations. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2018, page 1884. NIH Public Access.

Andrew K Lampinen, Ishita Dasgupta, Stephanie CY Chan, Kory Matthewson, Michael Henry Tessler, Antonia Creswell, James L McClelland, Jane X Wang, and Felix Hill. 2022. Can language models learn from explanations in context? *arXiv preprint arXiv:2204.02329*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Frederick Liu and Besim Avci. 2019. Incorporating priors with feature attribution on text classification. *arXiv preprint arXiv:1906.08286*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2021a. Noisy channel language model prompting for few-shot text classification. *arXiv preprint arXiv:2108.04106*.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2021b. Metaicl: Learning to learn in context. *arXiv preprint arXiv:2110.15943*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

Shikhar Murty, Pang Wei Koh, and Percy Liang. 2020. Expbert: Representation engineering with natural language explanations. *arXiv preprint arXiv:2005.01932*.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. 2021. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. *arXiv preprint arXiv:1906.02361*.

Shashank Srivastava, Igor Labutov, and Tom Mitchell. 2017. Joint concept learning and semantic parsing from natural language explanations. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 1527–1536.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Albert Webson and Ellie Pavlick. 2021. Do prompt-based models really understand the meaning of their prompts? *arXiv preprint arXiv:2109.01247*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot in-context learning. *arXiv preprint arXiv:2205.03401*.

Eric Zelikman, Jesse Mu, Noah D Goodman, and Yuhuai Tony Wu. 2022. Star: Self-taught reasoner bootstrapping reasoning with reasoning.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.