Review of the article: Can language models learn from explanations in context?

Andrew K. Lampinen, Ishita Dasgupta, Stephanie C. Y. Chan
Kory Mathewson, Michael Henry Tessler, Antonia Creswell
James L. McClelland, Jane X. Wang, Felix Hill
DeepMind
London, UK

A. *Introduction*

The article was written by a group of researchers from DeepMind and submitted to EMNLP 2022. It builds upon the topic of few-shot prompting for large language models to improve performance on new tasks. The authors investigate whether providing explanations for few-shot examples can further enhance the performance of LMs. This motivation stems from the fact for humans, explanations that connect examples to task principles can improve learning.

B. *Motivation and Research Question*

Few-shot prompting of Language Models has emerged as a new paradigm in natural language processing, where large LMs exhibit in-context learning abilities. They can perform a new language task by inferring from a few input-output pairs within the context window, without extensive training. As the authors mentioned, the ability to adapt with a few hints is similar to human flexibility with instructions or examples. However, explanations also play a central role in human learning. They can clarify the intended task by illustrating the principles that link questions to answers. Therefore, the authors were interested in investigating whether LMs can also benefit from explanations when learning from examples in-context, in order to generalize broadly.

The main research question of the article is:

Can explanations of answers improve few-shot task performance?

C. *Related Works*

This work is not the first to explore explanations. Given the increasing interest in prompt tuning, there has been other work on how in-context auxiliary information can affect model performance (e.g. Wei et al., 2022; Reynolds and McDonell, 2021). In this study, however, the authors include a particular focus on the effect of post-answer explanations, whereas other studies (e.g. Wei et al., 2022) provide chains of reasoning before the answers. The authors claim that pre- and post-answer explanations have different effects on model reasoning and evaluation, and offer distinct scientific insights. Furthermore, the proposed ideas in this work were evaluated on a distinctly challenging and diverse task set

D. *Main Contributions*

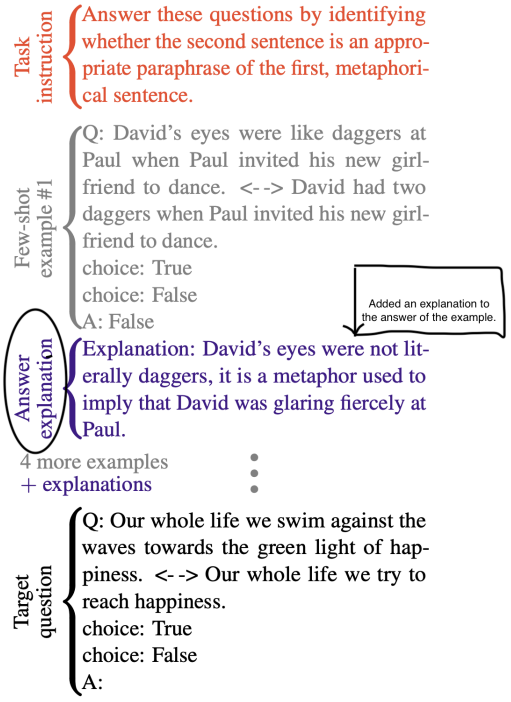
- Were annotated 40 diverse, challenging language tasks with explanations of examples, and released these annotations (see Fig. 1).
- Were evaluated several LMs after prompting with or without few-shot examples, explanations, instructions, and control conditions.
- It has been shown that explanations of examples in a few-shot prompt can improve the performance of large models. Even without tuning, they outperform matched control conditions.
- It has been shown that explanations tuned or selected using a small validation set can have larger effects.

E. *Approaches*

1. Dataset:

- A subset of 40 tasks and subtasks were selected from the BIG-Bench tasks datasets, which span a variety of reasoning types, skills, and domains.

Fig. 1: Example prompt including a task instruction, few-shot examples with explanations, and the target question. Task instructions outline the task to be performed, before the examples (or alone, in a zero-shot prompt). Answer explanations can be added to each of the examples in a few shot prompt. Performance is always evaluated on a target question. Because explanations are added to examples on a separate line after answers, the same evaluation metric can be used for the target question regardless of whether explanations are provided.



- Adding explanations to a prompt alters other prompt features, such as total length and content. To identify whether one of these lower-level features drives the effect of explanations, the authors crafted the following control explanations that match various aspects of the semantics, word- or sentence-level content:

  Scrambled explanations: To ensure that benefits are not due to word-level features, the authors compared to explanations with the words shuffled.

  True non-explanations: To test that it is the explanatory content that matters, the authors compared to a valid, relevant, but non-explanatory statement.

  Other item explanation: Finally, the authors evaluated whether the benefits were due to the direct relationship between the explanation and the explanandum, rather than some other feature of the language. To do so, they took the examples in a few-shot prompt and permuted the explanations so that the explanations did not match the question or answer, but the overall set

2. Model:

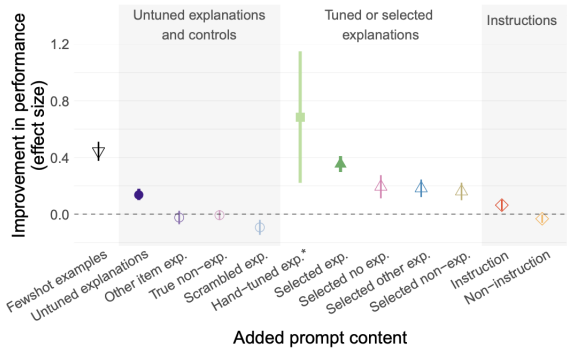
- For evaluation, the authors were provided with a set of LMs (Rae et al., 2021) ranging from 1 billion to 280 billion parameters

Evaluation:

- For each task, the authors constructed several 0- and 5-shot prompts (see Fig.1) in a subset of the possible combinations of task instructions (none, instruction, non-instruction) and explanations of examples (none, other item explanation, true nonexplanation, scrambled explanation).
- The model performance was evaluated in each prompt condition on all task dataset items (except those included in the prompt). The authors were restricted to multiple-choice tasks and evaluated the model's likelihood of each answer option after conditioning on the prompt and question (Fig. 1). They did not normalize the likelihoods by answer length, but they claimed that answer lengths are generally similar within a question, and in some preliminary experiments, such normalization did not improve performance. Answers were chosen greedily, i.e., the highest likelihood from the set, and the model's accuracy was scored according to the answer scores defined by the task, which may allow multiple correct answers or partial credit.
- Evaluating the models under different conditions yielded a complex dataset of results with hierarchical, nested dependencies. To estimate more precisely the effect of different components of the prompt on performance, the authors used hierarchical/multilevel modeling methods. Specifically, the authors fitted hierarchical logistic regressions that accounted for the multiple nested and heterogeneous structure of the results. These models considered dependencies from overall task difficulty, question idiosyncrasy, and prompt content. The models also accounted for diverse prompt effects on different tasks. Each factor was assessed through a specific parameter.

F. Results

Fig. 2: The benefits of different components of a prompt for the largest LM (280B), as estimated from hierarchical logistic regression. Each point estimates the unique added contribution of that component of the prompt. (Error bars are model 95%-CIs. Effect size estimated as log odds ratio / 1.81, see Chinn, 2000. *Explanations were only hand-tuned on five challenging tasks, so the CI is larger.)



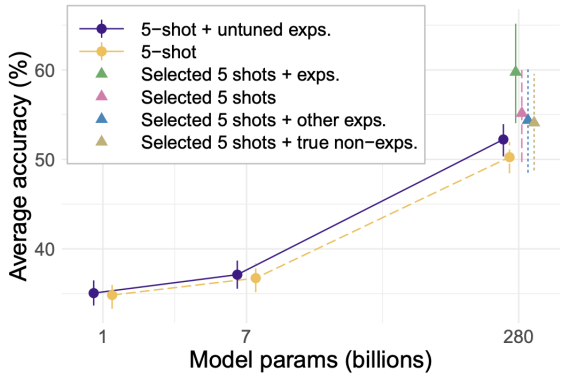
Summary of results from Fig.2:

- Adding few-shot examples to the prompt substantially improves performance relative to a zero-shot prompt and adding explanations (solid points) further improves performance.
- Even untuned explanations help, and tuned or selected explanations have substantial effects.
- Control conditions are neutral or harmful.
- Instructions provide a small benefit.

Summary of results from Fig.3:

- Untuned explanations lead to modest increases in accuracy from the largest model relative to few-shot prompts without explanations.

Fig. 3: The effects of explanations on average accuracy, across model sizes. **Note:** that these results aggregate across all 40 tasks, which have different difficulties and numbers of possible answers



- Smaller models do not benefit. A hierarchical regression confirms that larger models benefit significantly more from explanations
- Selecting the few-shot examples with explanations using a small validation set improves substantially over selecting only the few-shot examples without explanations, or with control explanations.

G. Summary of the Review

- **Contributions:**
  - The authors investigated a new strategy of few-shot promoting for challenging language tasks. They showed that explanation of the examples in few-shot prompting can increase the performance of the large LM's in some tasks.
  - They annotated 40 diverse, challenging language tasks with explanations of examples, and published these annotations.
- **Strengths**
  - The contributions of this work support the claims of the authors.
  - A clear research question and motivation make sense.
- **Weakness**
  - The authors could provide a clearer explanation of the specific types of explanations used in their method, and how they are generated. Additionally, it would be helpful to provide more details about the process for selecting and pre-processing the explanations.
  - The authors could consider including a more detailed analysis of the differences between their method and other related work (g. Wei et al., 2022) that has explored explaining reasoning before the answer
  - Although the dataset has been released, the article lacks detailed information about the tasks in the dataset. Some sample assignments could give a clearer image of the topic.

Overall, the reviewed paper provides valuable insights into the topic of prompting and can be considered a good article.