Few-shot prompting with large language models

Lorenc Zhuka

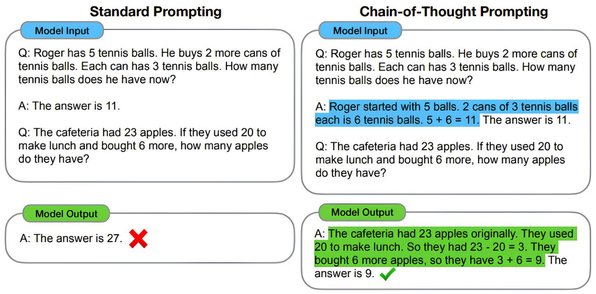
Darmstadt University of Applied Sciences - Data Science

Motivation

Large Language Models (LLMs) are deep learning models that are pre-trained on a large amount of unlabeled data and fine-tuned for downstream tasks. However, fine-tuning requires large labeled datasets and computational resources. Few-shot prompting (also known as in-context learning) which is a training free method, adresses these challenges. It involves the utilitization of a small set of examples to construct a context that serves as task demonstration, and this context is combined with a query or a question to create a prompt. This is fed into a pre-trained LLM and therefore no parameter update is conducted. Based on the provided prompt, the LLM generates an answer by extracting insights from the given examples. However, the effectiveness of few-shot prompting methods significantly depends on the choice of in-context demonstrations. Various techniques, such as Auto-CoT [3], Active-CoT [1], and Retrieval-CoT [3] have demonstrated their effectiveness, emphasizing diversity (Auto-CoT), uncertainty (Active-CoT), and similarity with the test example (Retrieval-CoT). Notably, while some methods, like Random-CoT and Auto-CoT, utilize labels generated by a LLM, Active-CoT relies on human-generated labels.

Chain-of-thought prompting

Standard prompting is the standard way of prompting a LLM which involves the construction of a few examples that demonstrate a specific task, which includes the question and the final answer, whereas chain-of-thought prompting involves generating a chain of thought, which is a series of intermediate reasoning steps that leads to the final answer [2]. Chain-of-thought prompting has been demonstrated to outperform standard prompting on tasks that require reasoning.



Objective and Research Questions

This thesis aims to evaluate and compare several prompting methods using closed-source and open-source models in arithmetic reasoning tasks under both labeling scenarios: labels generated by GPT-3.5-turbo with Zero-Shot-CoT and human-generated labels. It also aims to develop new prompting methods by combining diversity, uncertainty, or similarity with the test example.

- How can we effectively combine diversity and uncertainty to identify and select the most informative examples from a dataset?
- How do various few-shot prompting methods perform on reasoning tasks compared to Random baseline under different labeling scenarios , and how does the choice of labeling affect their few-shot performance?
- Can few-shot prompting with GPT-3.5-turbo outperform GPT-4 in zero-shot settings for reasoning tasks?
- How do open-source models such as Falcon perform on reasoning tasks when employed in few-shot prompting scenarios and how does the model size influence the performance?

Existing Prompting methods

- **Random-CoT** selects a few examples randomly.
- **Retrieval-CoT** selects semantically similar examples of a test example based on cosine similarity. As an LLM may generate incorrect answers when labeling, in Retrieval-CoT, it can lead to "misleading by similarity", where the LLM may replicate similar mistakes when reasoning for the test question.
- **Diverse-CoT** may mitigate the issue of "misleading by similarity" through the selection of diverse examples by partitioning them into a few clusters and sampling an example from each cluster. Examples closer to the cluster centroid are prioritized. Figure 1 illustrates a visual example. [3] call it Auto-CoT.
- **Active-CoT** selects examples that exhibit the highest LLM uncertainty. It has been demonstrated that reducing the LLM uncertainty improves the few-shot performance.

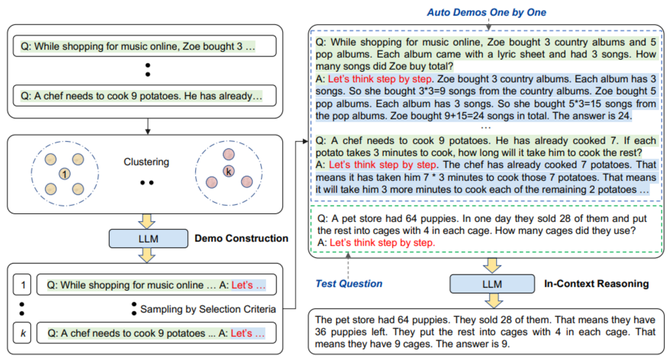


Figure 1. Example of Auto-CoT. Reprinted from [3]

Proposed Prompting Methods

- **Diverse-Active-KMeans-CoT** combines the strategies from Diverse-CoT and Active-CoT by partitioning the questions into a few clusters and selecting the example with the highest LLM uncertainty.
- **Diverse-Active-KMeansPlusPlus-CoT** prioritizes diversity and uncertainty in example selection using a weighted F1-score metric. It begins with the example having the highest uncertainty and iteratively selects the next examples based on a weighted F1-score, considering the similarity between questions and previously chosen examples.
- **Diverse-Active-KMeansPlusPlus-Retrieval-CoT** utilizes the Diverse-Active-KMeansPlusPlus-CoT method in the first phase to select p examples. In the second phase, Retrieval-CoT further refines the selection by choosing the most similar examples from the initially selected p examples.

Selection criteria of simple examples

GPT-3.5-turbo with Zero-Shot-CoT is one method of labeling questions. However, especially with complex examples, it may lead to inaccuracies due to the long reasoning steps generated by the LLM, especially with high token-count questions. To address this, a selection criterion prioritizes simpler examples by setting limits on reasoning steps and question tokens.

Results

Comparison of prompting methods

- On the GSM8K dataset, Diverse-CoT, Retrieval-CoT, Active-CoT, Diverse-Active-KMeansPlusPlus-CoT, and Diverse-Active-KMeansPlusPlus-Retrieval-CoT each surpass Random-CoT by margins of 1.1%, 1.7%, 3.5%, 1.9%, and 2.5%, respectively.
- Similarly, the AQUA dataset evaluation showcases Diverse-CoT, Active-CoT, and Diverse-Active-KMeansPlusPlus-Retrieval-CoT outperforming the Random approach by margins of 3.9%, 2.7%, and 3.3%, respectively.
- These results hold when GPT-3.5-Turbo is utilized for labeling questions and when reasoning chains are included in the in-context demonstrations, as opposed to the use of standard prompting. Conversely, when human-generated labels are used or standard prompting, these methods achieve performance comparable to Random-CoT or even exhibit lower performance.

Comparison of LLMs

- Zero-Shot-CoT with GPT-4 consistently outperforms all few-shot chain-of-thought prompting methods implemented with GPT-3.5-turbo, surpassing them by a significant average margins under different scenarios, 6.7% on GSM8K and 10.5% on AQUA.
- Falcon-40B-Instruct achieves 37.3% accuracy on GSM8K and 18.5% on AQUA, while Falcon-7B-Instruct achieves 5.4% on GSM8K and 11.4% on AQUA, underscoring the superior performance of larger models in few-shot settings. However, Falcon-40B-Instruct and Falcon-7B-Instruct demonstrate inferior performance compared to GPT-3.5-turbo, with margins exceeding 30%.

Future Work

- Expanding research using larger training datasets can enhance understanding of few-shot performance, as it can increase the effectiveness of certain prompting methods, such as Active-CoT, which may benefit from a dataset that has examples, where the model is less confident.
- Future research could focus on fine-tuning open-source language models on datasets like GSM8K and AQUA, using innovative methods like Low-Rank Adaptation of Large Language Models (LoRA) and QLoRA (Quantized LoRA) to adapt these models to reasoning tasks and compare their performance to more advanced models like GPT-3.5-turbo.
- Future research should explore ensemble techniques, combining multiple few-shot prompting methods like Auto-CoT, Active-CoT or Retrieval-CoT. There is a need for developing a unified approach with fewer hyperparameters. This stands in contrast to the proposed Diverse-Active-KMeansPlusPlus-Retrieval-CoT, which, due to its optimization of the parameter p, presents a drawback that impacts few-shot performance.

References

[1] Diao, S., Wang, P., Lin, Y., and Zhang, T. (2023). Active prompting with chain-of-thought for large language models.

[2] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. (2023). Chain-of-thought prompting elicits reasoning in large language models.

[3] Zhang, Z., Zhang, A., Li, M., and Smola, A. (2022). Automatic chain of thought prompting in large language models.

lorenczhuka28@yahoo.com