Investigating the Impact of In-Context Learning on LLM Models

Mira Zilberstein, Omri Fahn, David Kleinman

ABSTRACT

This study investigates the impact of In-Context Learning (ICL) on the performance of two Large Language Models (LLMs), GPT-3.5-turbo and Llama-2-7b-chat-hf, in the domain of question answering. Utilizing the "StrategyQA" dataset, we explore how the models' performance and running time correlate with different numbers of source documents. Our analysis examines the influence of external sources on the accuracy of the LLMs, challenging the prevalent notion that an increased contextual depth invariably leads to improved model performance. Traditionally, it's understood that LLMs can gain from in-context learning and an increase in the number of source documents, up to a certain threshold. However, contrary to our initial hypothesis and prevailing studies, our project did not record a marked improvement with a mere increase in the number of source documents as we will see. The findings highlight the complex interplay between the number of source documents and the internal mechanisms of LLMs in generating responses. This study contributes to the broader discourse on optimizing LLMs for complex Natural Language Processing tasks, offering insights into the nuanced dynamics of in-context learning and its implications for model development and application.

1 INTRODUCTION

In recent years, the field of Natural Language Processing has undergone a transformative shift with the introduction of Large Language Models. These models have showcased remarkable proficiency in generating contextually relevant text, solidifying their role as integral components in diverse NLP applications. One innovative approach that has gained prominence in harnessing the capabilities of LLMs is In-Context Learning (ICL).

ICL represents a transformative strategy that allows LLMs to adapt to new tasks without the need for explicit retraining. This ability is particularly crucial in real-world scenarios where the tasks at hand may evolve or change over time. The motivation behind utilizing In-Context Learning lies in its potential to strike a balance between achieving superior results and maintaining computational efficiency.

The primary objective of the current project is to investigate the influence of In-Context Learning on the performance of two selected LLMs, GPT-3.5-turbo and Llama-2-7b-chat-hf, for question answering task using the “StrategyQA” dataset. Specifically, the investigation delves into how variations in model performance and running time correlate with different numbers of source documents. To execute this exploration effectively, the project leverages cutting-edge platforms, namely Hugging Face and Langchain.

Subsequent sections of the paper will delve into the intricate details of the study, presenting experimental results, discussing observed patterns and drawing insightful conclusions based on the findings.

2 RELATED WORK

In this section, we discuss a range of projects focused on the impact of in-context learning on the performance of Large Language Models (LLMs). This examination is vital in contextualizing the main findings of our research, which will be further detailed in upcoming sections.

A study which investigates a different perspective to the effect of ICL is "Complementary Explanations for Effective In-Context Learning"[3]. This study underscores the significance of the quality and diversity of explanations in the learning process, rather than just focusing on the sheer number of source documents. It suggests that LLMs gain more from a varied set of explanations, showcasing different reasoning skills, rather than just an increased volume of documents. The study introduces a novel strategy for selecting exemplars based on maximal marginal relevance, aiming to balance relevance with diversity and thereby enhance in-context learning.

This understanding aligns with the broader narrative in the field, as evidenced by other significant works like “Leveraging Large Language Models for Multiple Choice Question Answering”[1] and “Learning to Retrieve In-Context Examples for Large Language Models”[2], which also emphasize the nuanced aspects of LLM training and application. Such a realization hints that our emphasis on increasing the quantity of source documents could have been complemented by a greater focus on enhancing their quality, potentially leading to more effective outcomes.

3 METHODOLOGY

In this section, we present our methodology for investigating the impact of In-Context Learning (ICL) on Large Language Models (LLMs), focusing on GPT-3.5 and Llama-2-7b-chat-hf. The methodology encompasses model setup, data preparation, implementation of the RetrievalQA chain and evaluation.

3.1 Model Setup and Configuration

We utilized GPT-3.5 from OpenAI and Llama-2-7b-chat-hf from Hugging Face. For Llama-2-7b-chat-hf, we configured the model with torch data types like `float16` and established a text generation pipeline, optimizing settings for sampling and token generation.

3.2 Data Collection and Preprocessing

The "StrategyQA" dataset was used. We collected the data using Python requests and split it into training and testing subsets. Training data was processed into Document objects for embedding and retrieval.

3.3 Embedding and Retrieval Mechanism

We employed FAISS for efficient similarity searching of dense vectors. Training set documents were embedded using 'sentence-transformers/all-MiniLM-L6-v2' and a FAISS index was created.

3.4 Implementation of RetrievalQA Chain

Two RetrievalQA chain instances were set up, one for each LLM. This chain retrieves relevant documents and generates responses. The FAISS database served as the retriever and we used a specific prompt template for response generation.

3.5 The "StrategyQA" Dataset

The "StrategyQA" dataset is central to our study, characterized by its focus on strategic question-answering. This dataset challenges LLMs with questions that require inference and reasoning, making it suitable for evaluating the in-context learning capabilities of the models.

3.5.1 Data Preparation. We accessed the "StrategyQA" dataset through a web request, dividing it into training and testing sets. The training set was further processed into Document objects for embedding and retrieval, aligning with our methodology's needs.

3.5.2 Relevance to Study. The complexity of the "StrategyQA" dataset, demanding contextual understanding and strategic reasoning, aligns with the objectives of In-Context Learning. It serves as an effective tool for assessing the adaptability and reasoning depth of the LLMs under investigation.

4 RESULTS

Our research aimed at evaluating the performance of two Large Language Models (LLMs) - GPT-3.5 and Llama-2-7b-chat-hf - in the context of In-Context Learning (ICL), particularly focusing on question-answering tasks using the "StrategyQA" dataset. The results, derived from a comprehensive analysis of model outputs, provide several key insights into the capabilities and limitations of these models.

4.1 Model Accuracy

- **GPT-3.5:** This model exhibited a higher accuracy, with a success rate of approximately 60.59%. This suggests a robust understanding and processing capability within the GPT-3.5 framework for the given dataset.
- **Llama-2-7b-chat-hf:** The Llama model registered a lower accuracy of around 45.43%. This disparity in performance compared to GPT-3.5 could be attributed to differences in their training data, model architecture, or other inherent characteristics.

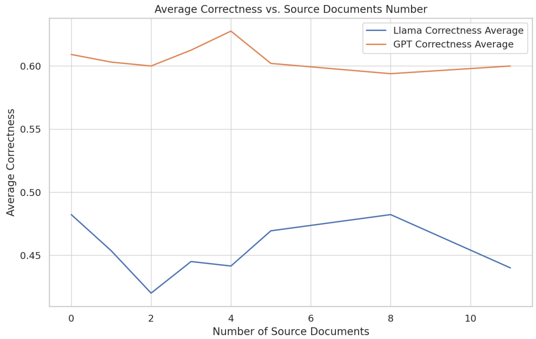

Figure 1: Average correctness of GPT and Llama models across different numbers of source documents.

4.2 Concurrent Model Inaccuracy

- **Rate of Both Models Being Incorrect:** In about 10.04% of the instances, both models failed to provide the correct answer. This could highlight the complexity of certain questions or a common gap in the knowledge base or reasoning capabilities of both models.

4.3 Distribution and Influence of Source Documents

- **Similar Distribution Across Models:** The distribution of the number of source documents was found to be similar for both models, suggesting a comparable approach in utilizing external information.
- **Mean Documents When Correct vs. Incorrect:** For GPT, the mean was 4.25 (correct) vs. 4.30 (incorrect) and for Llama, it was 4.27 for both. This further indicates that the number of source documents used is not a significant factor in determining correctness.

4.4 Run Time Average

- **Platform Considerations:** Although GPT exhibits lower run times than Llama, this comparison is influenced by the models running on different platforms. Direct comparisons cannot be made as open AI models run only on their servers.
- **No Correlation with Source Documents:** Our analysis shows no significant correlation between run time and the number of source documents for both models.

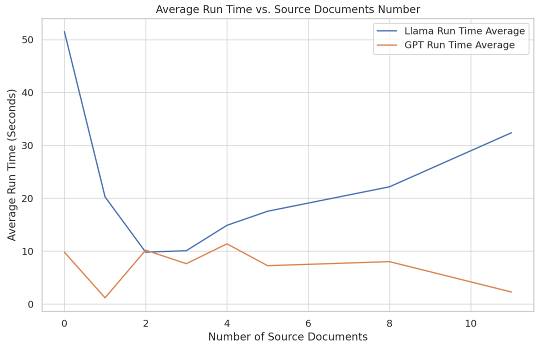

Figure 2: Average run time comparison between GPT and Llama models.

4.5 Answer Distribution

- **GPT-3.5:** This model tended to answer 'false' more frequently than 'true', indicating a possible inclination towards negation in uncertain scenarios.
- **Llama-2-7b-chat-hf:** In contrast, the Llama model predominantly answered 'true', suggesting a different approach or bias in processing and responding to queries.

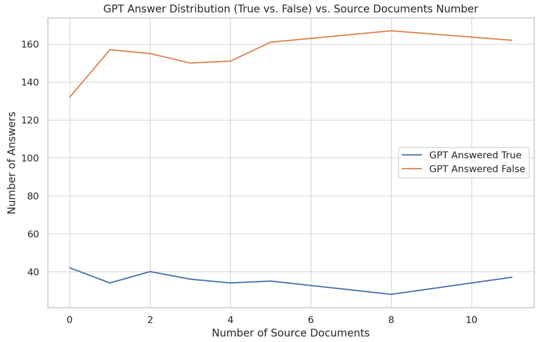

Figure 3: GPT model's answer distribution between 'true' and 'false' answers.

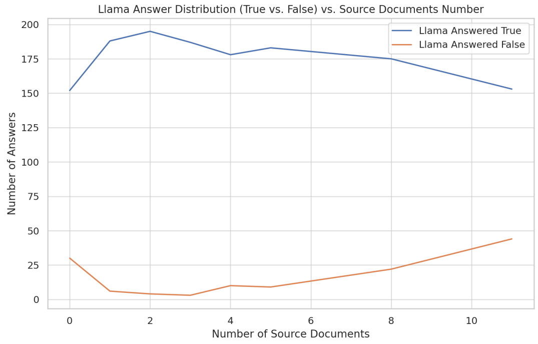

Figure 4: Llama model's answer distribution between 'true' and 'false' answers.

4.6 Correlation with Correctness and Answer Type

- **Correlation with Correctness:** Both models showed almost no relationship between the number of source documents and the correctness of answers (GPT: -0.0069, Llama: -0.0004).
- **Correlation with Answer Type:** For the GPT model, the correlation between the number of source documents and the type of answer (True, False, Other) is -0.1002. For the Llama model, this correlation is -0.1540. These negative correlations suggest that as the number of source documents increases, the likelihood of the models providing a definitive answer (True or False) rather than "Other" decreases slightly.

5 DISCUSSION

In this section, we critically analyze and interpret the findings of our study, exploring their implications and the broader context within the field of Natural Language Processing (NLP) and In-Context Learning (ICL).

5.1 Interpretation of Results

Our study revealed several key insights about the performance of GPT-3.5 and Llama-2-7b-chat-hf models in question-answering tasks. Notably, the GPT-3.5 model demonstrated superior accuracy compared to Llama-2-7b-chat-hf. This might be attributed to differences in model architectures, training datasets, or inherent processing capabilities. The higher accuracy of GPT-3.5 suggests that it may be better suited for tasks requiring nuanced understanding and contextual processing.

5.2 Role of Source Documents

Contrary to our initial hypothesis, the quantity of source documents utilized by the models did not significantly correlate with the correctness of their answers. This finding challenges the common assumption that a greater number of reference documents invariably leads to improved performance in LLMs. It raises questions about the efficiency and effectiveness of sourcing external documents for in-context learning, suggesting that the models' internal knowledge bases and reasoning algorithms play a more vital role.

5.3 Implications for Model Development

The similar patterns in the utilization of source documents by both models, regardless of their differing accuracies, provide important insights for future model development. It suggests that enhancing the quality and relevance of source documents might be more beneficial than merely increasing their quantity. Moreover, the occurrence of concurrent inaccuracies in both models for certain questions points to potential areas of improvement in their training or knowledge acquisition processes. This notion aligns with related works mentioned in this paper, which emphasize the importance of the quality and diversity of explanations in learning. These works point to a similar need for focusing on the richness rather than the quantity of context provided to the models.

5.4 Future Research Directions

The insights gained from this study open several avenues for future research. Investigating the impact of source document quality, diversity and relevance on model performance could yield valuable findings. Additionally, exploring other datasets and model architectures could help in understanding the generalizability of our results and furthering the development of more robust LLMs.

5.5 Limitations of the Study

While our study provides significant contributions to the understanding of in-context learning in LLMs, it is important to acknowledge its limitations. The specificity of the "StrategyQA" dataset and the choice of only two models might affect the generalizability of our findings. Future studies could address these limitations by incorporating a wider range of datasets and LLMs.

Also, it's important to recognize the resource constraints we encountered. For instance, the number of source documents we could utilize was limited, a factor that likely impacted the breadth of our research.

In conclusion, our study sheds light on the intricate dynamics of in-context learning in LLMs. The findings emphasize the need for a nuanced approach to leveraging external sources and underscore the importance of internal model mechanisms in achieving high-quality responses. This study thus contributes to the ongoing discourse on optimizing LLMs for complex NLP tasks.

6 CONCLUSION

In summary, our investigation into the influence of in-context learning on GPT-3.5 and Llama-2-7b-chat-hf models in question answering has yielded insightful results. Contrary to expectations, we found no substantial correlation between the number of examples in context and the accuracy of model-generated responses. This challenges the prevailing belief that increased contextual depth inherently leads to improved performance.

While our findings contribute valuable insights, certain limitations, such as dataset specificity and model choice, should be considered. Future research could explore diverse datasets, model architectures and evaluation metrics to broaden our understanding.

This study prompts a reevaluation of assumptions about in-context learning, laying the groundwork for further research. Our results underscore the complexity of language model behavior and advocate for a comprehensive and nuanced approach in optimizing in-context learning for language models.

7 NOTE

Link to the code used to evaluate the models' accuracy:

https://github.com/davidkleinman/nlp_proj_tau

REFERENCES

[1] Joshua Robinson, Christopher Michael Rytting, and David Wingate. 2023. Leveraging Large Language Models for Multiple Choice Question Answering. (2023). arXiv:cs.CL/2210.12353

[2] Liang Wang, Nan Yang, and Furu Wei. 2023. Learning to Retrieve In-Context Examples for Large Language Models. (2023). arXiv:cs.CL/2307.07164

[3] Xi Ye, Srinivasan Iyer, Asli Celikyilmaz, Ves Stoyanov, Greg Durrett, and Ramakanth Pasunuru. 2023. Complementary Explanations for Effective In-Context Learning. (2023). arXiv:cs.CL/2211.13892