LLM Response Report

Report on the Accuracy of Large Language Models in Answering ConvFinQA Questions

This report analyzes the performance of two large language models, Google Gemini and OpenAI ChatGPT, in answering questions from the ConvFinQA dataset. The dataset consists of 81 questions related to financial topics, each paired with relevant text, tables, and a correct answer. Our goal is to compare the accuracy of these LLMs in providing correct answers to these financial questions.

1. Introduction

The ConvFinQA dataset is designed to assess the ability of language models to understand and interpret financial information presented in a conversational context. The questions require LLMs to process text, tables, and sometimes even figures, and extract relevant information to provide accurate answers.

This report focuses on analyzing the accuracy of Google Gemini and OpenAI ChatGPT in answering these questions and identifying potential shortcomings and areas for future improvement.

2. Methodology

Our analysis follows these steps:

1. **LLM-Driven Prototype:** We utilized the provided ConvFinQA dataset to query both Google Gemini and OpenAI ChatGPT. Each question, along with its associated text and table information, was fed into the respective LLM.

2. **Answer Comparison:** The LLM-generated answers were then compared against the correct answers provided in the dataset.

3. **Accuracy Calculation:** We calculated the accuracy of each LLM by determining the number of correctly answered questions and dividing it by the total number of questions in the dataset.

4. **Shortcoming Analysis:** We identified questions where the LLMs provided incorrect answers and analyzed the question details (text and tables) to understand potential reasoning for the errors. This involved identifying the aspects of the question that might have posed challenges for the LLMs.

3. Results

LLM	Correct Answers	Incorrect Answers	Accuracy
Google Gemini	51	48	63%
OpenAI ChatGPT	61	39	75%

OpenAI ChatGPT outperformed Google Gemini in answering the ConvFinQA questions, achieving an accuracy of 75% compared to Gemini's 63% accuracy.

4. Findings

Our analysis revealed several interesting findings:

* **LLMs struggled with questions requiring calculations:** Both LLMs had difficulty performing

calculations, especially when the numbers involved percentages, ratios, or implied values. For instance, questions 16, 16b, 22, 29, 30, 35, 41b, 51b, 66, 68b, and 76b, all involved calculations, and both LLMs struggled to provide the correct answers.

* **LLMs struggled with questions asking for specific details from tables:** Many questions required LLMs to identify specific numbers from tables, which often involved recognizing relevant data points within a complex table structure. Questions 6, 9, 10, 11b, 17, 17b, 20, 24, 25, 26, 27, 37, 40, 41, 42, 42b, 44, 45, 49, 51, 52, 53, 54, 55, 58, 59, 61, 62, 64b, 68, 73, 76b, 79, and 81 are examples of such questions where LLMs struggled.

* **LLMs sometimes misunderstood the context of questions:** Certain questions, especially those requiring complex sentence structures or wordplay, were challenging for the LLMs to interpret correctly. This led to incorrect answers, despite the presence of accurate information in the provided text and tables. Questions like 5, 14, 32, 41, 42, 43, 58, 64, 77, and 80 are examples of questions where context misinterpretation played a role.

5. Shortcomings

The analysis identified several potential shortcomings of both LLMs:

* **Limited numerical reasoning ability:** Both LLMs showed limitations in performing calculations and deriving numerical insights from provided data.

* **Challenges with complex table structures:** LLMs struggled to efficiently extract information from complex tables, especially when questions required recognizing subtle relationships or patterns within the data.

* **Contextual understanding issues:** The LLMs demonstrated limitations in accurately interpreting the context of questions, particularly when questions involved nuanced language or complex sentence structures.

6. Improvements

To improve the accuracy of these LLMs in answering ConvFinQA questions, future work could focus on:

* **Enhancing numerical reasoning abilities:** Training the LLMs on a dataset that specifically emphasizes numerical reasoning and computation skills could improve their ability to handle complex calculations and derive quantitative insights from data.

* **Improving table comprehension:** Developing techniques that allow LLMs to more effectively parse and interpret complex table structures, including relationships and patterns, is crucial.

* **Enhancing contextual understanding:** Training the LLMs on a wider range of conversational and financial language patterns, with a focus on complex sentence structures and nuanced language, would enable them to interpret questions more accurately.

7. Conclusion

This report demonstrates that while LLMs like Google Gemini and OpenAI ChatGPT are capable of understanding and processing financial information, they still have significant limitations when it comes to providing accurate answers to complex financial questions. The LLMs, especially Gemini,

struggled with questions requiring numerical reasoning, understanding table structures, and accurately interpreting the context of questions. This highlights the need for continued research and development to improve the ability of LLMs to handle these specific challenges. By focusing on enhancing numerical reasoning abilities, improving table comprehension, and refining contextual understanding, we can expect to see greater accuracy and reliability in LLM responses to financial questions in the future.