FinTech Frankenstein Project Proposal

Akaash Dash
GA Institute of Technology
adash37@gatech.edu

Yahya Hassan
GA Institute of Technology
yhassan30@gatech.edu

Erik Larson
GA Institute of Technology
elarson43@gatech.edu

Aditya Natham
GA Institute of Technology
anatham3@gatech.edu

Sapan Patel
GA Institute of Technology
spatel794@gatech.edu

Max Zhao
GA Institute of Technology
qzhao305@gatech.edu

Abstract

In the rapidly evolving landscape of Language Model (LLM) development, the finance domain witnesses a constant influx of new models. This proliferation creates a challenge—how to discern the most suitable LLM for specific use cases. Different LLMs may yield varying responses based on content and phrasing, necessitating an in-depth assessment. Our research endeavors to address this challenge by comprehensively evaluating both general and financial-specific LLMs, identifying their strengths and weaknesses. Our ultimate aim is to construct a user-friendly web application capable of harnessing the capabilities of selected LLMs. Users will be able to input questions, and the application will provide detailed and precise responses. To achieve this, we will conduct a rigorous benchmarking process, exploring various LLMs and potential combinations. As the benchmarking parameters are yet to be standardized, this initial phase of our work is critical in ensuring that the LLMs integrated into the web app are optimized to serve diverse user needs in the finance sector.

1 Introduction

As new Large Language Models (LLMs) are being rapidly published, it is becoming harder to decipher which model is best for specific prompting tasks and use cases. Small variations in content and phrasing can result in vastly different responses across models, creating strong discrepancies. This is especially true in the financial domain, where the use of general purpose models is becoming as common as using domain specific models. General purpose models may have vastly different outputs than models created for the specific domain, creating further discrepancies.

2 Problem Statement

The potential for large discrepancies and wide range of models available make it very difficult to know which model to choose when given a use case. This becomes especially more true in the finance domain, where different tasks require very different abilities. To address the issue, we propose a survey of the landscape of existing and newly released LLMs in the general and finance domains, an evaluation of their strengths and weaknesses, and performance benchmarks for the models for various tasks and categories of prompts.

3 End Goal

Building on our proposed problem solution, we intend to build a web app that allows users to prompt a specific LLM or combination of multiple LLMs to answer relevant questions that can account for multiple use cases ranging from document analysis to general financial inquiries. Before building the app though, we need to experiment with the various LLMs available, along with combinations of LLMs, and benchmark their performance to determine the pros and cons of each LLM for certain usecases. This benchmarking process may become highly involved as there's little to no current standardized process for determining the effectiveness of an LLM, so we plan on placing much of our initial efforts on defining the parameters in which we're benchmarking around to ensure the final LLMs that are present in the web app is optimal for the users' use cases.

4 Literature Review

- https://arxiv.org/abs/2109.14394
  - EDGAR-CORPUS: Tokenized corpus of all SEC filings publicly available on EDGAR
- https://universal-ner.github.io/
  - Universal-NER: Large NER model trained on various data sources in a va-

riety of domains to provide LLM-level NER recognition

- https://arxiv.org/pdf/2302.11157.pdf
  - FiNER: Large data source for financial NER training/validation/testing data
- https://arxiv.org/abs/2303.17564
  - BloombergGPT: 50B LLM trained in the financial domain for a variety of financial tasks
- https://ai.meta.com/blog/large-language-model-llama-meta-ai/
  - LLaMA: Pretrained model that can be fine-tuned for financial domain.
- https://arxiv.org/abs/2306.06031
  - FinGPT: Large data source/framework for all tasks/models/related works on LLMs in the financial domain.
- https://ai.google/discover/palm2/
  - PaLM 2: Pretrained model that can be fine-tuned for financial domain.
- https://arxiv.org/pdf/2303.18223.pdf
  - Survey of LLMs: Large survey paper on the development and current state of LLMs, focus on open source
  - FinBERT: Adaptation of BERT for financial domain. Encompasses a range of tasks including sentiment analysis, NER, etc.

5 Proposed Methodology

5.1 Research

Research the various LLMs, as illustrated in Table 1 (at the top of last page), for pros and cons.

5.2 Test and Benchmark

Test and benchmark the previously researched LLM's:

1. Create a set of benchmarks covering a range of NLP tasks (NER, Sentiment, Prompt Generation, Prompt Response, etc.) in a financial context
2. Test each of the LLM's in a financial context and note each's performance
3. Benchmark the performances of each LLM and compare them to each other

5.3 Webpage

Webpage will include a list of various LLMs to choose from

1. Each LLM will have certain pros and cons highlighted which the user can understand for the given question
2. Text box the user can enter text in, and select which LLM they want to use

5.4 Web Page Expansion

Expand webpage with automatic model selection

1. Build a model on top of pro/con list to categorize the prompt input and select the appropriate model. Can use some sort of RLHF to do this.

Limitations

The main limitations of our project stems from the fact that our project aims to aggregate the leading LLMs and analyze them. This will inevitably cause a lot of expenditure. The API's for these LLMs are not free, nor are they very cheap. Because we will be benchmarking them on a multitude of tasks, and we will be allowing users to select which LLM they want to use for their prompt, we will be heavily relying on the APIs.

References

https://arxiv.org/abs/2109.14394
https://universal-ner.github.io/
https://arxiv.org/pdf/2302.11157.pdf
https://arxiv.org/abs/2303.17564
https://ai.meta.com/blog/large-language-model-llama-meta-ai/
https://arxiv.org/abs/2306.06031
https://ai.google/discover/palm2/
https://doi.org/10.1111/1911-3846.12832
https://arxiv.org/pdf/2303.18223.pdf

Model Name	General / Domain Specific	Month and Year Published
FiNER	Specific	Feb 2023
FinBERT	Specific	Sep 2022
PaLM	General	Apr 2022
PaLM 2	General	May 2023
FinGPT	Specific	Jun 2023
LLaMa	General	Feb 2023
LLaMa 2	General	Jul 2023
Bloomberg GPT	Specific	May 2023
GPT 3.5	General	Mar 2022
GPT 4	General	Mar 2023

Table 1: There is a clutter of LLMs, both general and finance domain-specific, created and improved upon monthly.