First Paper: Deficiency of Large Language Models in Finance: An Empirical Examination of Hallucination

Introduction:

- This paper goes into depth about the challenges of a large language model in finance. It specifically goes into the hallucinations that they generate which have plausible but incorrect information.
- Currently, LLMs are changing how people are looking into fields of education, finance, and more. The topic we will be focusing on is finance and in this area, we have helpful LLM models such as FinBERT, BloombergGPT, and FinGPT which can help people with financial decisions. The issue is that they often produce hallucinations which lead to significant risks, such as financial losses for the user. This may lead to people losing trust in these financial LLMs.
- The two major challenges to fixing these issues are measuring hallucinations accurately and ensuring the LLMs perform reliably in the financial task given to them.
- This is a summary of what the paper will be going into.
    - In this paper, taking an empirical approach, we examine the hallucination behaviors of LLMs in financial tasks.
        - Examining Financial Knowledge: Empirical investigation into the LLMs" ability to retain and use financial knowledge.
        - Case Study: Given an analysis of LLM's performance in querying historical stock prices to evaluate their practical utility.
        - Mitigation Methods: We will assess the LLMs using four methods to reduce hallucinations and improve factual accuracy.
            - Few-shot prompting
                - Few-shot prompting is a technique where you provide a machine learning model, particularly a language model, with a small set of examples to guide its behavior for a specific task. Unlike traditional machine learning methods that require extensive training data, few-shot prompting allows the model to understand the task with just a handful of examples. This is incredibly useful when you have limited data or need quick results.
            - DoLa (Decoding by Contrasting Layers)
                - Our approach obtains the next-token distribution by contrasting the differences in logits obtained from projecting the later layers versus earlier layers to the vocabulary space, exploiting the fact that factual

knowledge in an LLMs has generally been shown to be localized to particular transformer layers. We find that this Decoding by Contrasting Layers (DoLa) approach can better surface factual knowledge and reduce the generation of incorrect facts. DoLa consistently improves the truthfulness across multiple-choice tasks and open-ended generation tasks.

- RAG (Retrieval Augmentation Generation)
  - Retrieval-augmented generation (RAG) is the process of optimizing the output of a large language model, so it references an authoritative knowledge base outside of its training data sources before generating a response. Large Language Models (LLMs) are trained on vast volumes of data and use billions of parameters to generate original output for tasks like answering questions, translating languages, and completing sentences. RAG extends the already powerful capabilities of LLMs to specific domains or an organization's internal knowledge base, all without the need to retrain the model.
- A prompt-based tool learning method that generates correct function calls

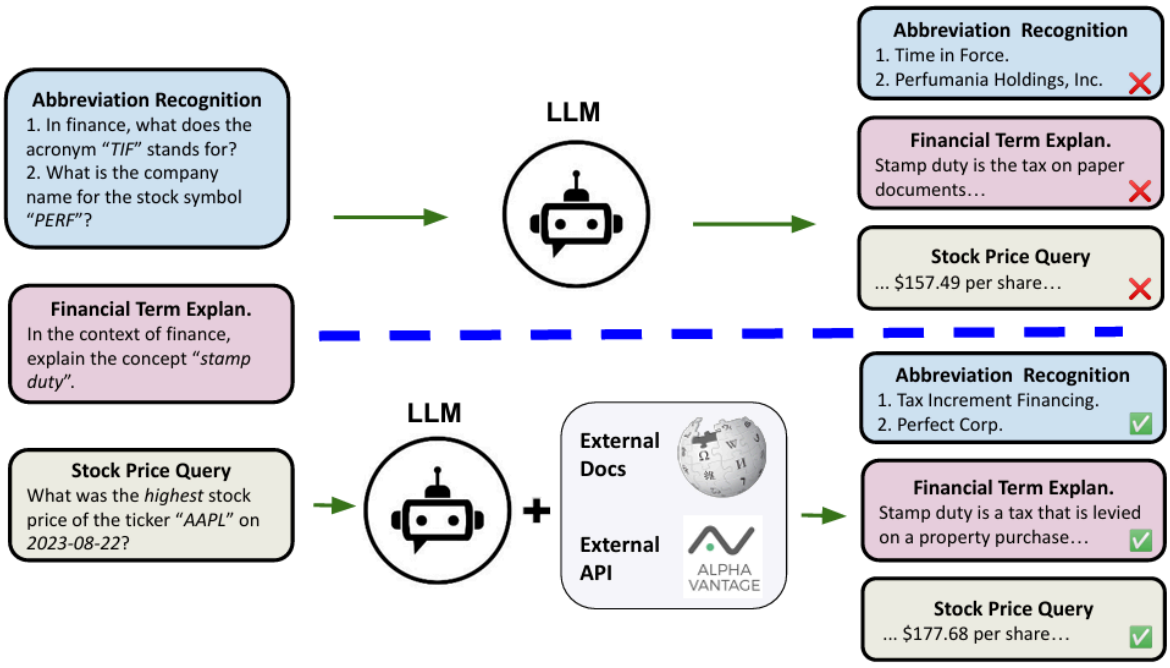
- In doing this they are hoping to enhance the reliability of LLMs for real-world financial applications.

Background and related works:

- When a user asks a question to an LLM, they expect a response that correctly correlates with the topic behind the user's question. However, the LLM may hallucinate their answer leading to an answer that may seem plausible but isn't correct. The GPT4 model for example misinterpreted the financial meaning of the acronym TIF, responding with "Time in Force" while the correct meaning of this acronym is "Tax Increment Financing".
- The hallucinations that the LLMs produce fall into four categories:
  - **Instruction inconsistency** - The LLM ignores the specific instructions given by the user.
    - For example, instead of translating a question into Spanish as instructed, the model provides the answer in English.
  - **Input context inconsistency** - The model output includes information not present in the provided context or contradicting it.

        - For example, an LLM claimed the Nile originates from the mountains instead of the Great Lakes region, as mentioned in the user's input.
    - **Generated context conflict** - The model output includes information from previously generated text that isn't correct.
    - **Factual inconsistency** - The model output includes information that isn't factually consistent.
        - This topic is particularly being focused on in this paper since it represents a serious and frequent type of error.
- RAG is one way to help with LLM hallucinations. This technique uses external knowledge sources and API calls which allow the LLM to get information from proper sources. This paper uses the DoLa method (Decoding by Contrasting layers) and a prompt-based tool-learning method to mitigate hallucinations.
- DoLa - a traditional RAG approach that contacts differences in logits obtained from projecting the later layers versus the earlier layers to the vocabulary space. Exploiting the idea that factual knowledge in LLM has generally been shown to be in particular transformer layers.
- Areas such as finance, medicine and law are affected by LLM's increased chance of hallucinations.

Methodology for Empirical Examination:



-
    - An empirical framework for evaluating hallucination behaviors of LLMs within the financial field.

    - In this image, we see that an LLM on its own may get its information from anywhere leading to old and incorrect information. However we can see performance improves when the LLM takes in integrated external data sources.
- In the experiments conducted to get information for this data, they used HuggingFace weights of pretrained Llama2-7B model and instruction-tund+RLHF version Llama2-7B-chat. Utilized OpenAI API to use models of GPT3.5-turbo and GPT4. Used FinMA-7B-NLP which is a fin-tuned LLaMA-1-7B model.
- There are 3 financial tasks which are used to provide an assessment of LLM performance:
    - Task 1: Financial Abbreviation Recognition
        - The task is to measure the ability to recognize financial acronyms and stock symbols.
    - Task 2: Financial Term Explanations
        - The task is to ask the LLMs for an explanation of financial terminology.
    - Task 3: Stock Price Query
        - The task is to ask the LLMs for a historical stock price