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.

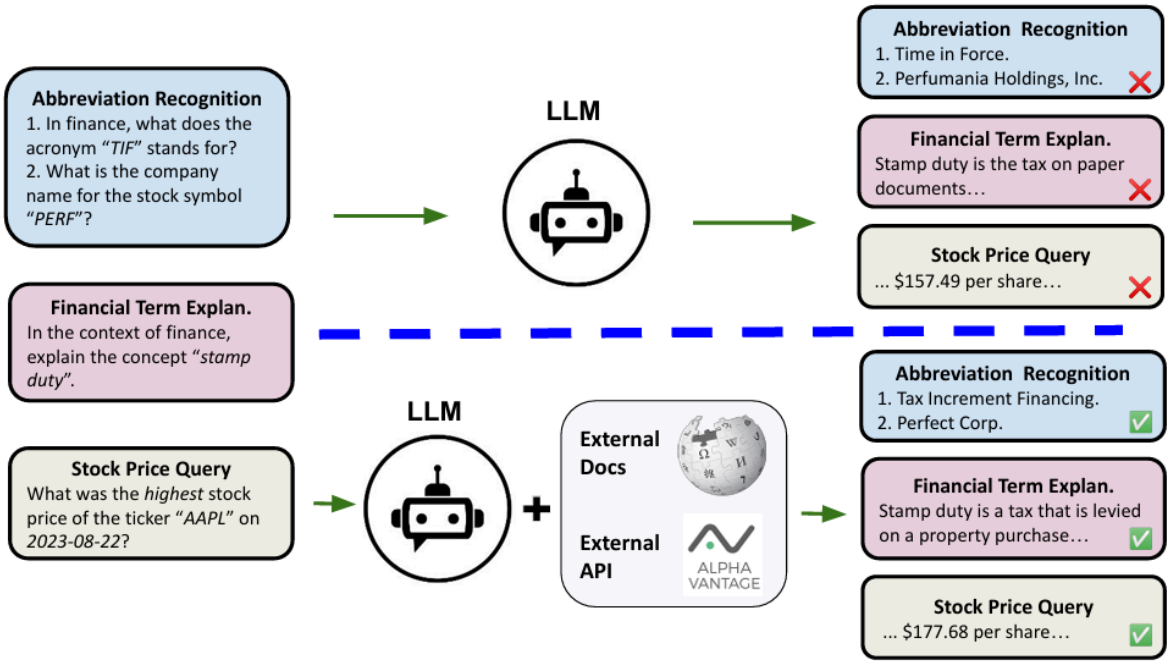
    - **Factual inconsistency** - The model output includes information that isn't factually consistent.
        - This topic is particularly being focused on in this paper since it represents a serious and frequent type of error.

- 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.

Empirical Results for Hallucination

Table 1: Results of the acronym recognition task and terminology explanation task. The GPT-3.5 Turbo and GPT-4 models abstain from answering 56.5% and 42.3% of questions, respectively. Their accuracy are evaluated only on the subset of questions to which they provide answers.

Model	Method	Acronym Accuracy (%)	Stock Symbol Accuracy (%)	Term explanation FactScore (%)
Llama1-7B	Zero-Shot	40.5	14.3	46.34
Llama2-7B	Zero-Shot	50.1	16.0	51.94
Llama2-7B-chat	Zero-Shot	74.8	16.5	64.68
Llama2-13B	Zero-Shot	58.3	24.2	56.08
Llama2-13B-chat	Zero-Shot	75.0	12.2	66.72
GPT3.5-turbo	Zero-Shot	76.8	87.7*	78.54
GPT4	Zero-Shot	82.5	90.4*	81.11
FinMA-7B	Zero-Shot	30.4	7.4	25.56
Llama1-7B	Few-Shot	54.3	22.7	49.27
Llama2-7B	Few-Shot	57.3	23.0	51.60
Llama2-7B-chat	Few-Shot	77.1	20.0	67.02
Llama2-13B	Few-Shot	62.7	25.4	63.28
Llama2-13B-chat	Few-Shot	76.1	23.7	67.97
FinMA-7B	Few-Shot	36.5	9.8	27.64
Llama2-7B	DoLa	41.5	17.3	56.39
Llama2-7B-chat	DoLa	61.8	17.6	65.42
Llama2-13B	DoLa	42.8	23.4	67.06
Llama2-13B-chat	DoLa	62.4	13.7	67.50
Llama2-7B	RAG	86.6	61.5	75.07
Llama2-7B-chat	RAG	90.8	**69.3**	90.48
Llama2-13B	RAG	87.4	62.8	78.56
Llama2-13B-chat	RAG	**93.4**	68.5	**90.67**

Table 2: The comparative performance of different models for the stock price query task, with and without an external tool. *The OpenAI models abstain to answer any question related to stock prices.

Models	Valid (%) ↑	MAE ($) ↓	int_correct (%) ↑	Accuracy (%) ↑
		Zero-Shot		
Llama2-7B	85.3	6357.6	2.0	0.0
Llama2-7B-chat	100.0	6380.5	2.7	0.0
GPT3.5-turbo*	0.0	-	-	-
GPT4*	0.0	-	-	-
Llama2-7B+DoLA	76.8	6857.9	0.9	0.0
Llama2-7B-chat+DoLA	97.4	6460.2	1.2	0.0
Llama2-7B+tool	47.5	-	-	76.4
Llama2-7B-chat+tool	89.8	-	-	91.2
GPT3.5-turbo+tool	100.0	-	-	98.4
GPT4+tool	100.0	-	-	100.0
		One-Shot		
Llama2-7B	90.2	489.3	2.1	0.0
Llama2-7B-chat	100.0	234.5	3.4	0.0
Llama2-7B+DoLA	86.6	521.6	2.1	0.0
Llama2-7B-chat+DoLA	97.4	234.9	2.4	0.0
Llama2-7B+tool	97.4	-	-	100.0
Llama2-7B-chat+tool	99.8	-	-	100.0
GPT3.5-turbo+tool	100.0	-	-	100.0
GPT4+tool	100.0	-	-	100.0

- General-purpose LLMs generate factually incorrect content in finance.
    - After benchmarking smaller open-source models, it was clear that it was hard for them to compile and use information in the financial area. Some of the incorrect responses provided by the LLM models were outdated leading to information that was obsolete.

- Multi-task domain-specific finetuning could diminish LMs' general instruction-following abilities.
    - They found that models that were turned for specific financial tasks performed worse than their base model.
    - While it was important to have the model tuned for their certain area of expertise, it may lead to a decrease in accurately following instructions and adapting to new tasks.

- General-purpose LLMs generate seriously unreliable real-world financial predictions.
    - When predicting stock prices, Zero-shot LLMs had high Mean Absolute Errors and low accuracy. This is also the same case for the few prompting models but they were closer to providing correct answers. The high deviations in answers suggest substantial inaccuracies in their predictions.
    - GPT3.5-turbo and GPT4 were more cautious, not responding with an answer to stock price-related questions and some stock symbol recognition questions due to a lack of external tools.
        - This is a good thing since they won't be spreading misinformation on the matter.

- RAG significantly improves the factuality in finance.
    - In Table 1, integrating RAG elevated the performance with the Llama-2 and Llama-2 models.
        - Specifically, they scored higher in FactScores in long-form generation tasks compared to other models.
    - This displays the importance of using RAG for pre-trained and instruction-tuned models.

- Prompt-based tool learning helps significantly with the time-sensitive task.
    - With the LLMs that use the prompt-based learning tool such as the Llama2-7B, they were able to achieve 100% accuracy given stock price queries.
    - Integrating zero-shot and few-shot tool learning is an effective strategy for bridging the knowledge gap and enhancing the reliability of LLMs.

- Few-shot learning better improves the ability to follow the question-answering format than factuality.
    - LLMs under the few-shot learning are significantly better in question-answering compared to LLMs under zero-shot learning.
    - However while few-shot learning eps in adapting to question formats, it isn't as helpful in factual precision.

- DoLa has limitations in enhancing models with knowledge gaps in training data.
    - This method is designed to allow LLMs to have better factual accuracy by contrasting outputs from the different layers of the model.
    - Assumes higher layers of the model contain more factual knowledge. (Reduce hallucinations and improve accuracy of responses with this)
    - However, its comprehensive knowledge is lacking in its pretraining dataset. Since it relies on this knowledge, its ability to compensate for gaps is constrained.

Conclusion, Limitations and Future Works

- To mitigate hallucinations, we show the effectiveness of the RAG method and prompt-based tool learning to generate correct function calls, providing up-to-date information.

- The tasks that this paper put up do not encompass the full spectrum of real-world tasks in the domain of finance.

Second Paper: A Survey of Large Language Models in Finance (FinLLMs)

Introduction:

- Researchers found that scaling models to sufficient sizes not only enhances model capacity but also allows for in-context learning that doesn’t appear in small-scale language models.

- The rapid advancement of general-domain LLMs led to an investigation into Financial LLMs using Mixed-domain LLMs with prompt engineering and Fine-tuned LLMs with prompt engineering.

- The paper focuses on LMs in English and here are the key contributions:
    - Exploring the evolution from general-domain LMs to financial-domain LMs

- Comparing five techniques across four financial PLMs (Pre-trained Language Models) and four financial LLMs, including training methods and data, and instruction fine-tuning methods.
- Summarize the performance evaluation of siz benchmark tasks and datasets between different models, and provide eight advanced financial NLP tasks and datasets for the development of advanced FinLLMs.
- Discuss opportunities and challenges of Fin-LLMs regarding datasets, techniques, evaluation, implementation, and real-world applications.

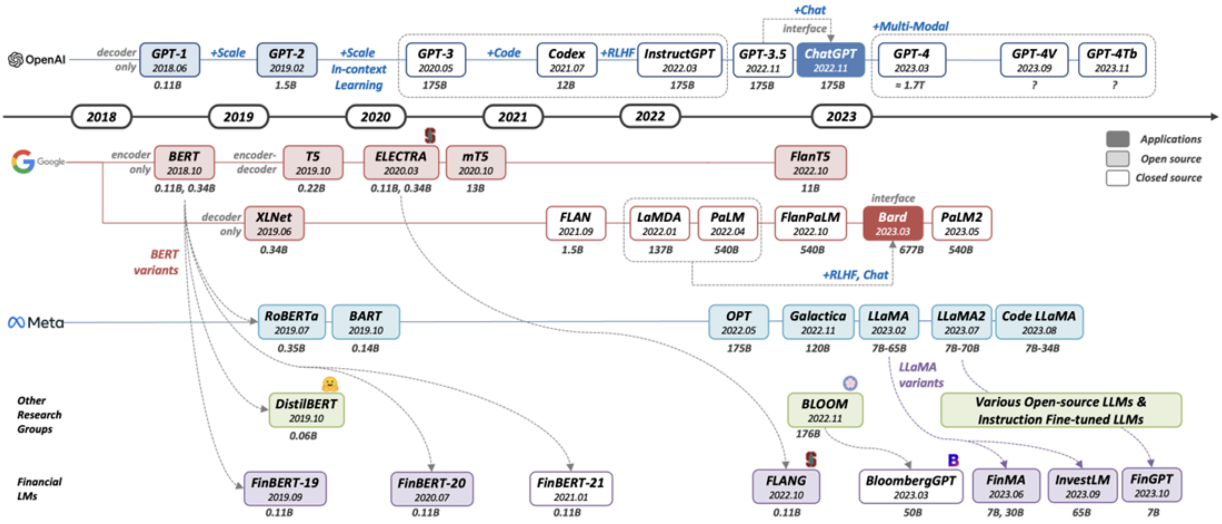


Figure 1: Timeline showing the evolution of selected PLM/LLM releases from the general domain to the financial domain.

Methodology for Empirical Examination:

- LLMs are generally trained with either discriminative or generative objectives.
  - Discriminative pre-training uses a masked language model to predict the next sentence and features and encoder-only or encoder-decoder architecture.
    - Encoder-only - uses only the encoder process to process the input sequence.
    - Encoder-decoder - uses both the encoder and decoder part of the transformer architecture to process the input sequence and processes the output.

  - GPT Series:
    - The first of the Generative Pre-trained Transformers (GPT) was GPT-1.
    - The next model GPT-2 was a more scaled model and used a probabilistic approach for multi-task problem-solving.
    - GPT-3 introduced in-context learning.

        - In-context learning - the model acquires capabilities that were not explicitly trained allowing language models to understand human language and produce outcomes beyond original pre-training objectives.
    - ChatGPT combined the in-context learning from GPT-3, LLM's for code from Codex (interprets commands in natural language and executes them on the user's behalf.), and reinforcement learning with human feedback RLHF from InstructGPT.
    - Success of ChatGPT lead to further development of models with GPT-4 demonstrating human performance capable of passing law and medical exams.
- Open-source LLMs
    - Before LLMs, we had PLMs such as BERT which was the foundational model for other PLMs. Then after OpenAI became closed-souce, there was a reduction in open souce models.
    - LLaMA came out later on which was a open souce LLM came out introducing techniques such as Instruction Fine-Tuning (IFT) and Chain-of-Thought (CoT) Prompting, the amount of open souces LLMs increased.
- Financial domain LLMs:

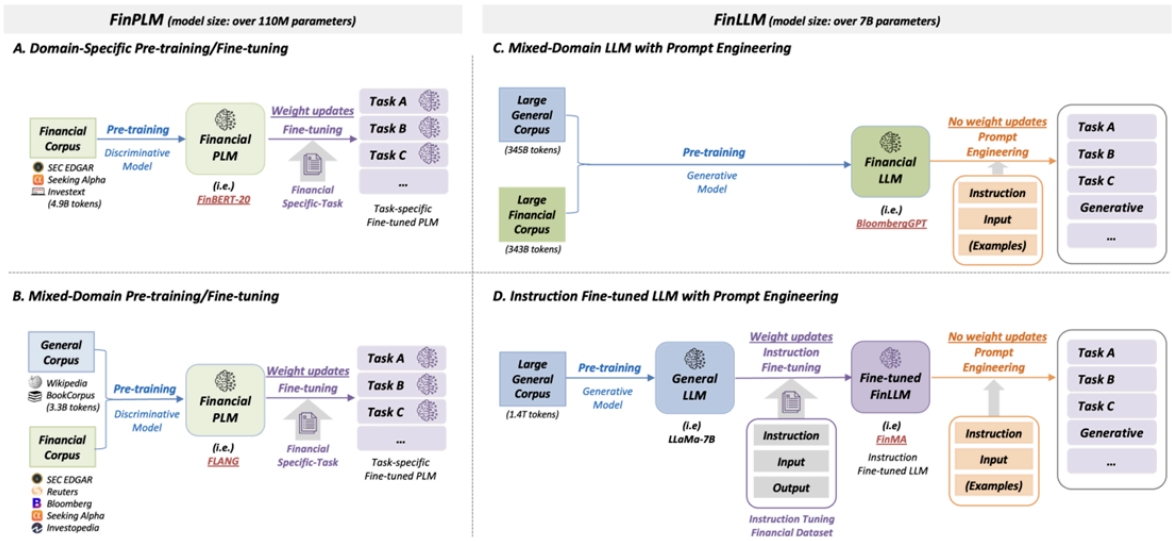


Figure 2: Comparison of techniques used in financial LMs: from FinPLMs to FinLLMs.

    - There are primary four financial PLMs and four financial LLMs.
        - PLMs:
            - FinPLMs:
                - FinBERT-19, FinBERT-20, and FinBERT-21 are based on BERT

        - FLANG is the fourth PLM and is based on ELECTRA
    - LLMs
        - FinLLMs:
            - FinMA, InvestLM, and FinGPT are based on on LLaMa and other open source based models
        - BLOOM-style closed source model
            - BloombergGPT

Techniques from FinPLMs to FinLLMs:

Category	Model	Backbone	Paras.	Techniques	PT	Evaluation		Open Source			Venue
					PT Data Size	Task	Dataset	Model	PT	IFT	
FinPLM (Disc.)	FinBERT-19 [Araci, 2019]	BERT	0.11B	Post-PT, FT	(G) 3.3B words (F) 29M words	[SA]	FPB, FiQA-SA	Y	N	N	ArXiv Aug 2019
	FinBERT-20 [Yang *et al.*, 2020]	BERT	0.11B	PT, FT	(F) 4.9B tokens	[SA]	FPB, FiQA-SA, AnalystTone	Y	Y	N	ArXiv Jul 2020
	FinBERT-21 [Liu *et al.*, 2021]	BERT	0.11B	PT, FT	(G) 3.3B words (F) 12B words	[SA], [QA] [SBD]	FPB, FiQA-SA, FiQA-QA FinSBD19	N	N	N	IJCAI (S) Jan 2021
	FLANG [Shah *et al.*, 2022]	ELECTRA	0.11B	PT, FT	(G) 3.3B words (F) 696k docs	[SA], [TC] [NER], [QA], [SBD]	FPB, FiQA-SA, Headline FIN, FiQA-QA, FinSBD21	Y	Y	N	EMNLP Oct 2022
FinLLM (Gen.)	BloombergGPT [Wu *et al.*, 2023]	BLOOM	50B	PT, PE	(G) 345B tokens (F) 363B tokens	[SA], [TC] [NER], [QA]	FPB, FiQA-SA, Headline FIN, ConvFinQA	N	N	N	ArXiv Mar 2023
	FinMA [Xie *et al.*, 2023]	LLaMA	7B, 30B	IFT, PE	(G) 1T tokens	[SA], [TC], [NER], [QA] [SMP]	FPB, FiQA-SA, Headline FIN, FinQA, ConvFinQA, StockNet, CIKM18, BigData22	Y	Y	Y	NIPS (D) Jun 2023
	InvestLM [Yang *et al.*, 2023c]	LLaMA	65B	IFT, PE PEFT	(G) 1.4T tokens	[SA], [TC] [QA], [Summ]	FPB, FiQA-SA, FOMC FinQA, ECTSum	Y	N	N	ArXiv Sep 2023
	FinGPT [Wang *et al.*, 2023]	6 open-source LLMs	7B	IFT, PE PEFT	(G) 2T tokens (e.g. LLaMA2)	[SA], [TC] [NER], [RE]	FPB, FiQA-SA, Headline FIN, FinRED	Y	Y	Y	NIPS (W) Oct 2023

Table 1: A Summary of FinPLMs and FinLLMs. The abbreviations correspond to Paras.= Model Parameter Size (Billions); Disc. = Discriminative, Gen. = Generative; Post-PT = Post-Pre-training, PT = Pre-training, FT = Fine-Tuning, PE = Prompt Engineering, IFT = Instruction Fine-Tuning, PEFT = Parameter Efficient Fine-Tuning; (G) = General domain, (F) = Financial domain; (in Evaluation) [SA] Sentiment Analysis, [TC] Text Classification, [SBD] Structure Boundary Detection, [NER] Named Entity Recognition, [QA] Question Answering, [SMP] Stock Movement Prediction, [Summ] Text Summarization, [RE] Relation Extraction; (in Venue) (S) = Special Track, (D) = Datasets and Benchmarks Track, (W) = Workshop. In open source, it is marked as Y if it is publicly accessible as of Dec 2023.

- Continual Pre-training:
    - Aims to train existing general LM with new domain specific data on incremental sequence of tasks.
    - An example of this is in FinBERT-19 which implemented 3 steps:
        - 1. Initialization of the general-domain BERT PLM
        - 2. Continual pre-training on financial domain corpus.
        - 3. Fine tuning on financial domain specific NLP tasks.
- Domain-Specific Pre-training from Scratch:
    - This approach involves training a model exclusively on an unlabeled domain-specific corpus while following the original architecture and training objective.
    - An example of this is FinBERT-20 which was pre-trained on financial communication corpus.
- Mixed-Domain Pre-training

    - This approach involves training a model using both a general-domain corpus and a domain specific corpus. The assumptio n is that general-domain tet remain relevant while the financial domain data provides knowledge and adaptation during the pre-training process.
    - An example of this is FinBERT-21 which employes multi-task learning across six supervised pre-trainng tasks allowing it to efficiently capture language knowledge and semantic information.

- Mixed-Domain LLM with Prompt Engineering:
    - Mixed-domain LLMs are trained on both a large general corpus and a large domain-specific corpus. Then, users describe the task and optionally provide a set of examples in human language. This technique is called Prompt Engineering and uses the same frozen LLM for several downstream tasks with no weight updates.
    - An example of the is BloombergGPT which uses the BLOOM model. It takes in a large general corpu and large financial corpus. The financial corpus, FinPile contains data from the web, news, filings, press, and Bloombers proprietary data.

- Instruction Fine-tuned LLM with Prompt Engineering
    - Instruction tuning is the additional training using explicit text instructions to enhance the capabilities and controllability of LLMs.
    - Research on instruction tuning can be classified into two main areas:
        - 1. The construction of instruction datasets
        - 2. The generation of finetuned LLMs using these instruction datasets.
    - An example is FinMA which is constructed from a large-scale multi-task instruction dataset called Financial Instruction Tuning. It also includes the Stock Movement Prediction task.
    - Other examples of this are InvestLM and FInGPT

Evaluation: Benchmark Tasks and Datasets:

- Sentiment Analysis:
    - The Sentiment Analysis task aims to analyze sentiment information from input text including financial news and microblog posts. Most FinPLMs and FinLLMs report the evaluation results of this task using the Financial PhraseBak and the FiQA SA dataset.
        - FPB is a dataset consisting of English financial news articles. Domain experts annotate each sentence in these articles and add one of three labels:

            - Positive,
            - Negative
            - Neutral
        - FiQA SA dataset consists of posts from headlines and microblogs.
            - They have sentiment scores on a scale of -1 to 1.
        - When it came to the task of Sentiment Analysis given the current PLMs and LLMs, it was found that PLM FLANG-ELECTRA achieved the best result with a 92% on F1 while the FinLLMs FinMA-30B and GPT-4 achieved an 87% on F1 using 5-shot prompting.
            - Since these result percentages were so similar, it suggest a practical approach for less complex tasks in terms of efficiency and costs.

- Text Classification (TC)
    - Text Classification - classifying a given text or document in predefined labels based on its content.
    - In financial text, there are often multiple dimensions of information beyond sentiment such as price directions or interest rate directions.
    - As TC is a broad task depending on the dataset and its predefined labels, three open-released financial TC datasets were added for further research.
        - FedNLP is a dataset comprised of documents sourced from various Federal Open Market Committee materials.
            - Annotated with labels Up, Maintain, Down based on the Federal Reserve's Federal Funds Tate decision for the subsequent period.
        - FOMC is a dataset collection of FOMC documents.
            - Annotated with labels as Dovish, Hawkish, or Neutral based on the prevailing sentiment conveyed within the FOMC materials.
        - Banking77 is a dataset comprised of samples covering intents related to banking customer service queries such as "card loss" or "linking to an existing card". This dataset is designed for intent detection and developing of conversation systems.

- Named Entity Recognition (NER)
    - NER - its task is the extraction of information from unstructured text and categorizing it into predefined entities such as locations (LOC), organizations (ORG), and persons (PER).

    - For the financial NER task, the FIN dataset is included in FLUE benchmarks.
        - The FIN dataset is comprised of eight financial loan agreements sourced by the US Security and Exchange Commission (SEC) for credit risk assessment.
    - For further research a financial NER dataset FiNER-139 was included consisting of sentences annotated with eXtensive Business Reporting Language (XBRL), word-level tags, sourced from the SEC
        - Designed for Entity Extraction and Numerical Reasoning tasks, predicting XBRL tags based on numeric input data within sentences
            - XBRL tags - an identifier for each individual item of financial data that are used to deliver human-readable financial statements in a machine-readable structure format
                - e.g. cash and cash equivalents
            - Numeric input data such as 24.8 million

- Question Answering (QA)
    - QA is a task to retrieve or generate answers to questions from an unstructured collection of documents.
    - Financial QA is more challenging than general QA since it requires numerical reasoning across multiple formats.
    - Over time, Financial QA has evolved to include complex numerical reasoning in multi-turn conversations.
        - This evolution involves the introduction of hybrid QA
            - Create alternate paths to connect hybrid contexts including tabular and textual content.
            - Examples of QA datasets
                - Financial QA: FiQA-QA
                - Hybrid QA: FinQA, ConvFinQA

- Stock Movement Prediction (SMP)
    - SMP - its task is it aims to predict the next day's price movement (price goes up or down) based on historical prices and associated text data.
    - It requires the integration of time series problems with temporal dependencies from text information, it presents a complex task where text data can act as both noise and signal.
        - FinMA includes SMP tasks
            - Examples of datasets used:
                - StockNet

        - Collected historical price data and Twitter data between 2014 and 2016 for 88 stocks listed in the S&P.
    - CIKM18
        - Utilizes historical price and Twitter data in 2017 for stocks in the S&P 500
    - BigData22
        - Compiled data between 2019 and 2020 for stocks in the US stock markets such as StockNet.
- Results from the tree datasets show accuracy of around 50%.
- It is important to consider financial evaluation metrics such as the Sharpe ratio as well as backtesting simulation resuslts

-