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

Introduction:

- This paper brings up the challenges of a large language model in finance. Specifically the hallucinations that they generate giving plausible but incorrect information.

- Currently, LLMs are changing how people look into 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. These LLMs help people learn about finance and making 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 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:
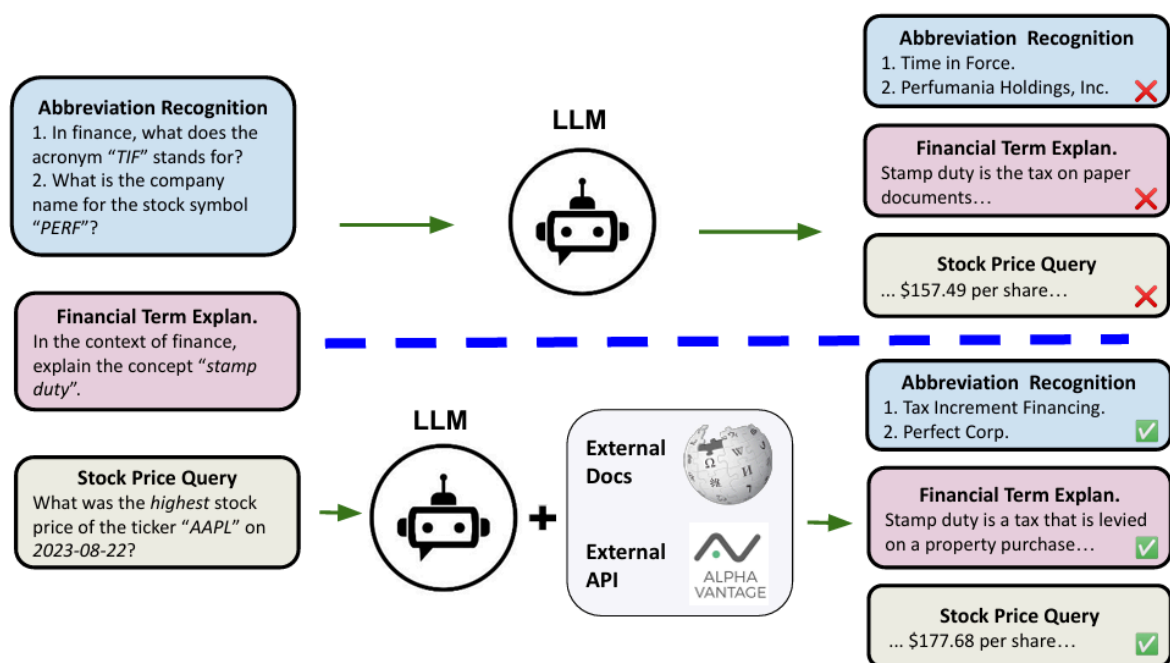
-
  - An empirical framework for evaluating hallucination behaviors of LLMs within the financial field.
-
  - In this image, we see that an LLM may get its information from anywhere leading to old and incorrect information. However, the performance improves when the LLM has integrated external data sources.

  - In the experiments conducted to get information for this data, they used HuggingFace weights of pre-trained Llama2-7B model and instruction-tund+RLHF version Llama2-7B-chat. They used OpenAI API to use models of GPT3.5-turbo and GPT4. They 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 difficult for LLMs 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 area of expertise, it could decrease its accuracy when 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. The few prompting models also had these problems but 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 as much.

- 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 than 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 LLMs that use prompt-based learning tools such as the Llama2-7B, it was 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 improves the ability to follow the question-answering format rather than factually.
    - 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 provide better factual accuracy by contrasting outputs from the different layers of the model.
    - This method 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 and provide 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 six benchmark tasks and datasets between different models, and provide eight advanced financial NLP tasks and datasets to develop advanced FinLLMs.

    - It will discuss the opportunities and challenges of Fin-LLMs regarding datasets, techniques, evaluation, implementation, and real-world applications.
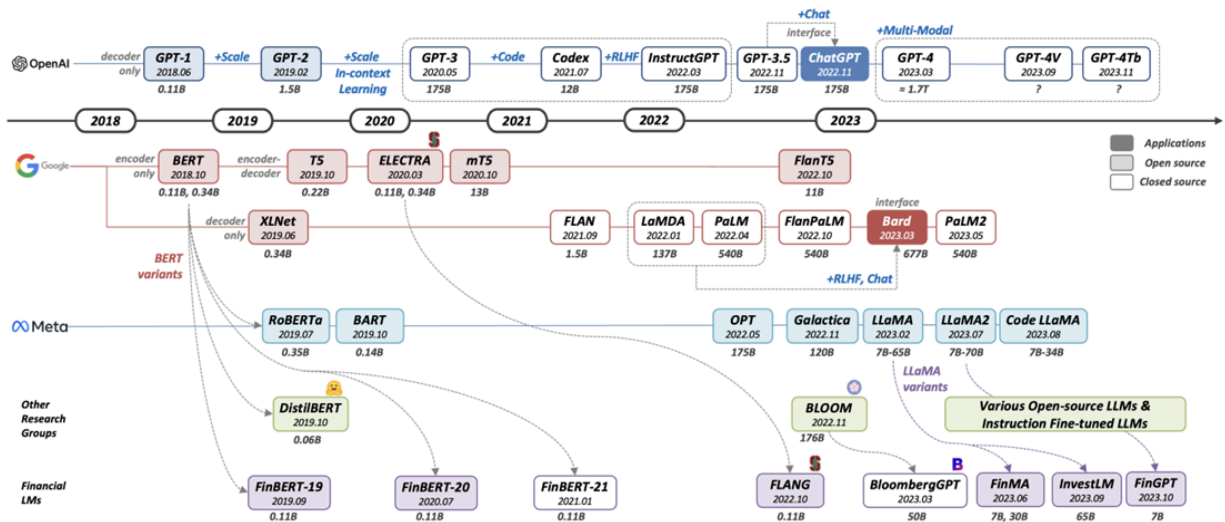
Figure 1: Timeline showing the evolution of selected PLM/LLM releases from the general domain to the financial domain.

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 process the output.

    - Generative pre-training will involve the model to predict the next word or sequence of words in a sentence giving the preceding words.
      - Here are some key concepts:
        - The model uses the context provided by the input text to generate relevant and coherent responses.
        - Generative Tasks: text completion, generation, translation, summarization, and more.
    - Difference:
      - **Discriminative Models:** Aim to classify or label data. They are trained to distinguish between different categories.

- **Generative Models:** Aim to generate new data. They are trained to produce text that is coherent and contextually appropriate.
- 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.
  - The success of ChatGPT led 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-source, there was a reduction in open-source models.
  - LLaMA came out later on which was an open-source LLM, introducing techniques such as Instruction Fine-Tuning (IFT) and Chain-of-Thought (CoT) Prompting, the amount of open-source LLMs increased.
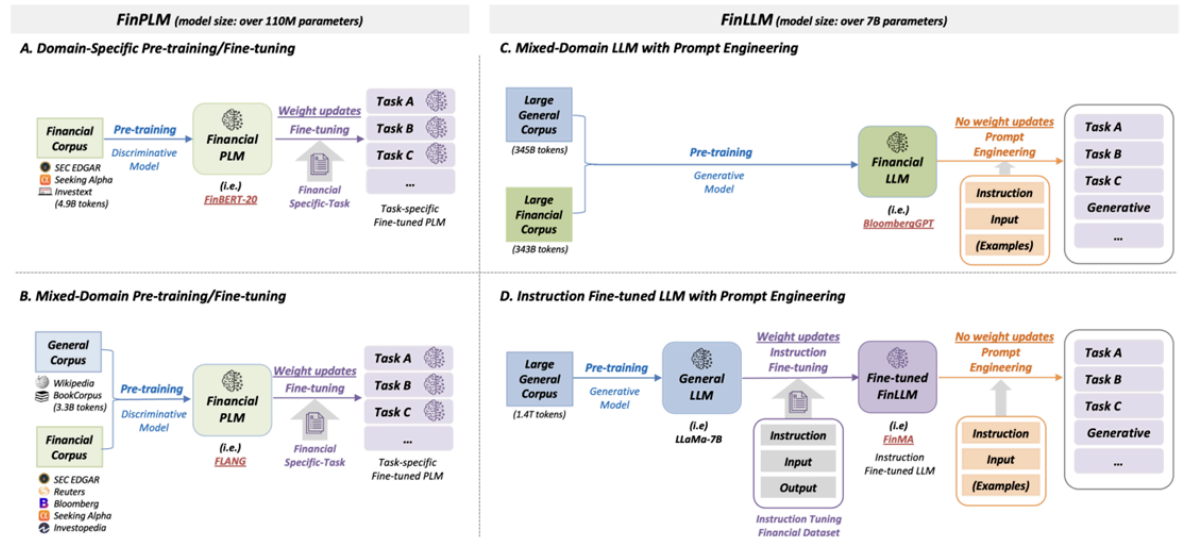
- Financial domain LLMs:

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 LLaMa and other open-source based models
    - BLOOM-style closed-source model
      - BloombergGPT

Techniques from FinPLMs to FinLLMs:

Category	Model	Backbone	Paras.	Techniques	PT Data Size	Task	Dataset	Model	PT	IFT	Venue
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:
  - This aims to train existing general LMs with new domain-specific data on the 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 assumption is that general-domain LM remains relevant while the financial domain data provides knowledge and adaptation during the pre-training process.

- ○ An example of this is FinBERT-21 which employs 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 domain-specific corpus. Then, users describe the task and optionally provide a set of examples in the 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 this is BloombergGPT which uses the BLOOM model. It takes in a large general corpus and a large financial corpus. The financial corpus, FinPile contains data from the web, news, filings, press, and Bloomberg 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 PhraseBank 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 suggests 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 the development 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 item of financial data that is 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 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 the 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 an accuracy of around 50%.

  - It is important to consider financial evaluation metrics such as the Sharpe ratio as well as the backtesting simulation results.

- Text Summarization (Summ)
  - Summ - the generation of concise summaries from documents while conveying key information using an extractive or abstract approach.

- Extractive
    - Picks the most relevant sentences from an article and systematically organizes them.

- Abstract
    - A more effective method:
        - Presents the material in a logical, well-organized, and grammatically sound form.

- Finance has been relatively underexplored due to the lack of benchmark datasets, challenges with domain experts' evaluations, and the need for disclaimers when presenting financial advice.

- Given the Financial LLMs and PLMs, they found that in this topic, they couldn't score higher than 50%

- The summarization task offers significant development opportunities, exploring whether FinLLMs can outperform task-specific SOTA models.

- Discussion
    - Mixed domain FinPLMs perform well in Sentiment Analysis (SA), Text Classification (TC), and Named Entity Recognition (NER) tasks. Fine-tuning a PLM for a specific task can be effective depending on task complexity.

- For Question Answering (QA), Semantic Similarity Prediction (SMP), and Summarization (Summ) tasks, task-specific state-of-the-art (SOTA) models outperform all large language models (LLMs), indicating room for improvement in FinLLMs.

- GPT-4 shows impressive performance across all benchmarks except for the Summ task, suggesting that simply scaling models may not be enough for optimal performance in finance.

- The inclusion of additional datasets for future research is recommended as most instruction fine-tuned FinLLMs used the same datasets for their evaluation.

Advanced Financial NLP Tasks and Datasets:
- For more complex financial tasks we have these 8 benchmark tasks

- - Relation Extraction (RE) - identifies and classifies relationships of entities based on the text.
    - FinRED dataset
      - financial news and earning call transcripts
    - The task is part of the Information Extraction

- (IN FINANCE) Event Detection (ED) - identifies the impact of how investors perceive and assess related companies.
  - Event-Driven Trading dataset
    - News articles with minute-level timestamps and stock price labels.

- (IN FINANCE) Causality Detection (CD) - seeks to identify cause-and-effect relationships within factual text and to develop the ability to generate meaningful financial narrative summaries.
  - FinCausal20 dataset
    - Shares two tasks
      - Detecting a causal scheme in a given text
      - Identifying cause-and-effect sentences.

- (IN FINANCE) Numerical Reasoning (NR) - Identifies numbers and mathematical operators in either digit or word form, This is so it can perform calculations and comprehend financial context (e.g. cash and cash equivalent)
  - Datasets designed for numerical reasoning
    - FiNER-139
    - FinQA
    - ConvFinQA

- Structure Recognition (SR) - focused on Structure Boundary Detection within a document (e.g. text, tables, or figures) and recognizing the logical relationships between tables and surrounding content, or between cells within a table.
  - FinTabNet dataset
    - Collected from earnings reports of S&P 500 companies

- Multimodal (MM) - integrates and processes data from multiple modalities such as text, audio, images, and video within a single framework.
  - MAEC dataset
    - It compiles the multimodal data (text, time series, and audio) from earnings call transcripts on a larger scale.
  - MONOPOLY dataset

- Introduces video data from monetary policy call transcripts across six central banks

- (IN FINANCE) Machine Translation (MT) - Translates sentences from a source language to a target language along with being able to comprehend the financial contextual meaning in different languages
  - MINDS-14 dataset
    - Contains samples of banking voice assistant data in text and audio formats across 14 different languages.
  - Mul-tiFin dataset
    - Contains samples covering financial topics with 6 high-level labels (e.g., Finance) and 23 low-level labels (e.g., M&A & Valuations), sourced from public financial articles in 15 different languages.

- Market Forecasting (MF) - involves the prediction of market price, volatility, and risk.

Opportunities and Challenges:
- Datasets: Developing advanced Financial Language Models (FinLLMs) requires high-quality, multimodal financial data. The challenge lies in collecting diverse, high-quality financial data and building instruction-fine-tuned datasets for specific financial NLP tasks.

- Techniques: Key challenges include using internal data without breaching privacy and enhancing trust in FinLLMs' responses. Techniques like Retrieval Augmented Generation (RAG) can enhance models by retrieving external knowledge to improve accuracy and minimize hallucinations while maintaining data privacy.

- Evaluation: Evaluating FinLLMs is challenging because it requires domain knowledge from financial experts. Standard NLP metrics like the F1-score are used, but human evaluation and specific financial metrics are essential for a comprehensive assessment.

- Implementation: The challenge in selecting suitable FinLLMs and techniques involves balancing cost and performance. Depending on task complexity, it may be more practical to use general-domain LLMs with specific prompts or task-specific models. This requires LLMOps engineering skills.

- Applications: Developing real-world financial applications faces non-technical challenges like business needs, industry barriers, data privacy, accountability, ethics, and bridging the understanding gap between financial and AI experts. Sharing FinLLM use cases across various financial fields is beneficial, and future directions include generative applications like report generation and document understanding.

Third Paper: A Survey of Large Language Models for Financial Applications: Progress, Prospects and Challenges

Introduction:
- In recent years, the financial domain has witnessed a growing interest in applying LLMs across various applications. These applications are not only reshaping the landscape of financial analysis but also offering new perspectives on market behavior and economic activities.
  - The tasks and details on how the LLMs perform:
    - Linguistic Tasks
      - Excel in summarizing and extracting key information from extensive financial documents,
        - Can change complex financial narratives into concise summaries which enable more efficient information processing.

- Sentiment Analysis
  - Can quantify market sentiment from financial news, social media, and corporate disclosures.
    - Allowing for critical insights that influence market movements and investment decisions.

- Financial Time Series Analysis
  - Can forecast market trends, detect anomalies, and classify financial data,
  - However, their power remains under debate.
    - Aims to enhance prediction accuracy and robustness by leveraging their deep learning architecture to capture complex temporal dependencies and patterns within financial datasets

- Financial Reasoning
  - Capable of reasoning

- It enables them not only to fit the data but also to emulate reasoning processes similar to human cognition

- LLMs support financial planning, generate investment recommendations, and assist in decision-making
    - It does this by processing and synthesizing vast amounts of financial data from diverse sources

- Agent-based Modeling
    - Extends the reasoning capabilities of LLMs to interactions between agents and their environments, markets, and humans
        - It can simulate market behaviors, economic activities, and the dynamics of financial ecosystems.

- They also present several challenges, such as lookahead bias in backtesting, legal concerns surrounding robot-generated content, data pollution, signal decay, inference speed, cost, uncertainty estimation, dimensionality considerations, interpretability, legal responsibility, safety, and privacy. Addressing these challenges is essential to ensure the ethical and effective deployment of LLMs in financial applications.
    - Robot-generated content - text, images, video, or audio content created by an artificial intelligence tool as a result of human input or prompts.

    - Data pollution - refers to the contamination of data with inaccurate, misleading, or irrelevant information. This can occur due to various reasons, such as errors in data entry, outdated information, or intentional manipulation.

    - Signal decay - refers to changes in data patterns and relationships that a machine-learning model has learned, potentially causing a decline in model quality.
        - This can lead to a divergence of decision boundaries for new data from those built from earlier data

    - Inference - the process that a trained machine learning model uses to conclude brand-new data.
        - Tests how well an AI model can apply information learned during training to make predictions or solve tasks

    - Cost - performance of the AI.

- Uncertainty estimation - the process of quantifying the uncertainty or confidence in the predictions made by an AI model.
  - Helps in understanding how reliable the model's predictions are in critical applications in our case finance.

- Dimensionality consideration - techniques used to reduce the number of variables or dimensions in complex datasets.

- Legal responsibility - protecting intellectual property ensuring data privacy and addressing liability for accidents or mishaps related to AI use.
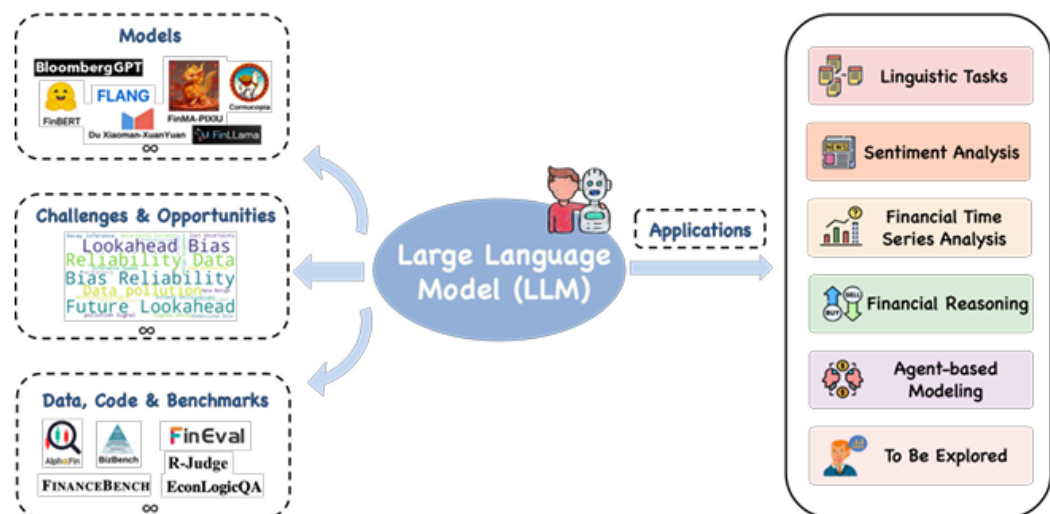

Fig. 1. An overview of our paper structure, focusing on models, applications, data, code and benchmarks, and challenges and opportunities.

-
- Collection of models:
  - GPT-series
    - Ploutos
      - LLM framework derived from GPT-4 proposed for interpertable stock movement prediction.
  - BERT
    - FinBERT-19
      - Developed by continually pre-traing BERT o financial text to enhance sentiment analysis capabilities.
    - FinBERT-20
      - Developed by conducting domain specific pre-training from scratch focusing solely on financial communicationsand utilizing a large-scalefinancail corpus.

- ■ FinBERT-21
  - ● Developed by mixed-domain pre-training leveraging both general corpora (Wikipedia and BooksCorpus) and f inancial domain corpora (FinancialWeb, YahooFinance, and RedditFinanceQA).
- ■ Mengzi-BERTbase-fin
  - ● Trained with 20GB of financial news and research papers
- ○ Text-to-Text Transfer Transformer a unified framework that treats every text processing task as a "text-to-text" problem.
  - ■ BBT (Big Bang Transformer)-FinTS
    - ● Incorporates knowlege enhanced pre-training methods and is built on a large-scale financial corpus including corporate reports, analyst reports, social media, and financial reports
      - ○ BBT-FinCorpurs
- ○ ELECTRA
  - ■ FLANG
    - ● Integrates specific adaptations such as selective token masking and span boundary objectives to effectively handle the complexities of financial language.
- ○ BLOOM
  - ■ BloombergGPT
    - ● Designed for financial sector by training on Bloomberg's financial data services.
  - ■ XuanYuan
    - ● Created for chinese financial market
    - ● Has hybrid-tuning strategy that combines general and financial-specific data allowing the model to retain generallanguage capabilities while excelling at domain-speific tasks such as financial advisory and market analysis.
- ○ Llama-series
  - ■ FinMA
  - ■ Fin-Llama
  - ■ Cornucopia - Chinese,
  - ■ Instruct-FinGPT and InvestLM
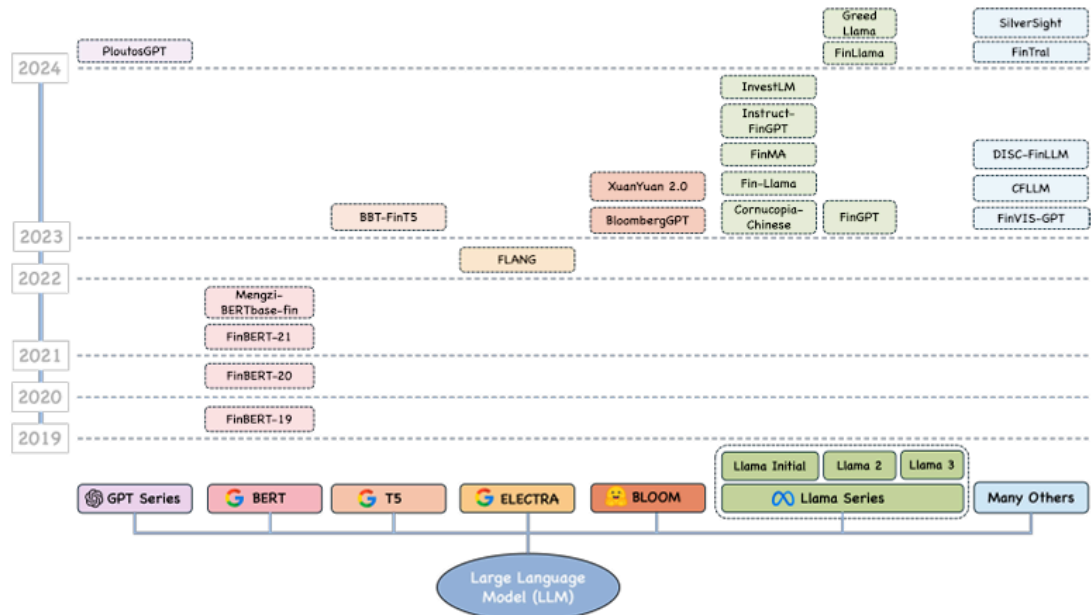 - ■ FinGPT
  - ■ FinLlama
  - ■ GreedLlama

Fig. 2. Overview of financially specialized large language models (LLMs) from 2019, categorized by their foundational model types and many others.

-
- Why Applying LLMs in Finance?
  - Core reasons for leveraging LLMs in financial applications:
    - Advanced Contextual Understanding:
      - LLMs can have a comprehensive understanding of financial terminologies, jargon, and refined expressions.
        - This enhances the accuracy of sentiment analysis which is important when dealing with complex language found in financial documents and news articles.

    - Transfer Learning Flexibility:
      - LLMs are trained using a large amount of internet text with a wide range of topics and languages.
        - This allows for LLMs to be fine-tuned for specific financial tasks.
        - The flexibility from transfer learning reduces reliance on large domain-specific datasets allowing for efficient adaptation to new tasks with minimal training on the specific topic data in finance.

    - Scalability for Real-time analysis.

- LLMs process large volumes of text rapidly and can come up with real-time reasoning and sentiment analysis.
  - This allows for financial decision-making from news articles, market information, reports, and social media.

- Multimodality:
  - This allows LLMs to read more than just text and include images, audio, and structured data to its information.
    - In finance, this could be text analysis from the news with visual analysis of stock price movements allowing for a more comprehensive understanding of market trends.

- Interpretability:
  - LLMs can generate human-like outputs.
    - This means they can create explanations for their results allowing us to understand how they came to that conclusion.
      - By providing results and underlying explanations, it builds trust and transparency with their data.

- Customization:
  - LLMs can accommodate different types of information which helps with the level of their analysis.
    - An example of this is integrating domain-specific data and parameters such as risk assessment for bonds or trend prediction in stock markets which the LLMs focus on
      - This will enhance the analysis generating insights that are tuned to the complexities of the different financial environments.

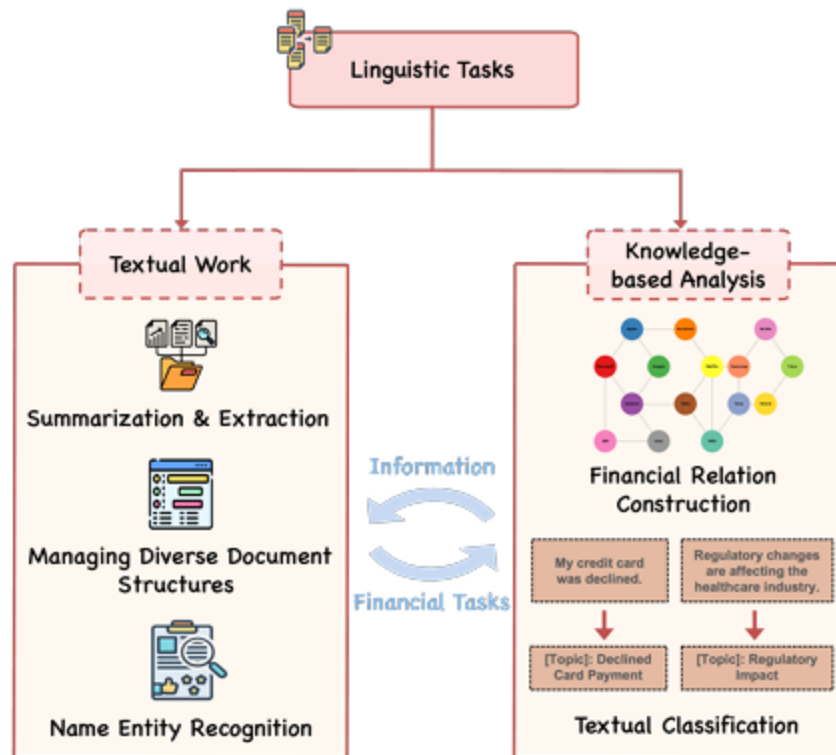**Chapter 3 Applications :**
- Linguistic Tasks:

Fig. 3. Illustration of various linguistic tasks in finance.

- ○
- ○ Models such as Recurrent Neural Networks and Long Short-Term Memory can understand text and perform textual work but struggle with long-term dependencies with long text sequences with complex expressions, large datasets, and unstructured data. This is critical in the Financial field.

- ○ LLMs can use the transformer architecture specifically its self-attention mechanism to process, understand, and generate text based on massive datasets they have been trained on making them work better than the two previous models mentioned

- ○ By efficiently managing long-term dependencies and contextual information over large amounts of text it can handle complex financial narratives and turn them into concise summaries that extract relevant information.
  - ■ Summarization and Extraction:
    - ● Financial documents are often lengthy leading to exceeded token limits of many LLMs.
      - ○ This is needed to divide documents into shorter segments to combat this.

- The idea to fix this was to use structural elements rather than paragraphs.
  - Segmenting long reports into ten distinct sections such as financial highlights, business overview, and more to aid the summarization process.

- Expanded to multilingual and domain-specific challenges
  - This allows for summarization in multiple languages and automates the fine-tuning process of text summarization without human annotation.

- Managing Diverse Document Structures:
  - LLMs sometimes face challenges with PDF documents that use images, charts, and tables.
    - One method to try to fix this was the Automated Financial Information Extraction framework which transforms tables into text using PLAIN serialization.
      - Allows for integrated table data with regular paragraphs to be processed by the LLM

    - This may lead to a loss of information embedded in charts or tables which was a problem.

- Name-Entity Recognition:
  - NER is necessary for extracting entities like company names, financial terms, and stock symbols from financial sources

  - Used for tasks such as industry classification, sentiment analysis, credit scoring, fraud detection, and regulatory compliance.

  - Traditional NER methods include:
    - Rule-Based Methods:
      - This method depends on hand-crafted rules.
      - This method has high precision but is not scalable.
    - Machine Learning Techniques:
      - This method includes supervised and unsupervised approaches
        - Supervised (HMM, Decision Trees, SVM)
    - Deep Learning Techniques:

- ○ This method uses architectures like BiLSTM, transformers, and CRF for better pattern recognition

    - ■ LLMs are enhancing NER by leveraging vast pre-trained knowledge

- ○ Knowledge-based Analysis:
    - ■ Financial Relation Construction:
        - ● Uses knowledge graphs to organize and analyze relationships between entities like companies, individuals, and financial instruments.
            - ○ Knowledge graphs help in predicting events and insights into them.

- ○ Question-Answering Systems:
    - ■ Knowledge graphs enhance multi-document questions to be answered by organizing information contextually.

- ○ Dynamic Knowledge Graphs:
    - ■ FinDKG is an example of a dynamic knowledge graph that adapts to market changes, aiding in strategic investment decisions.

- ○ Mask One At a Time (MOAT) Framework:
    - ■ Uses SEC-BERT to extract contextual embeddings and train neural networks for classifying relationships between financial entities.

- ○ In-Context Learning with GPT Models:
    - ■ This uses learning-free dense retrievers and learning-based retrievers for classifying financial relationships.

- ○ CFERE Framework:
    - ■ Focuses on multip-type Chinese financial event relation extraction using a combination of Core Verb Chain, Syntactic Semantic Dependency Parsing graph, and enhanced BERT

- ○ Textual Classification:
    - ■ This is used for organizing and understanding large volumes of unstructured data in the financial domain

    - ■ Company or Industry Classification:

- Grouping companies based on characteristics such as business activities and market performance.

- Traditionally relies on systems like GICS, SIC, NAICS, and the Fama French model.

- Could use LLMs to generate company embeddings from SEC filings.
  - Achieved accurate reproduction of GICS classifications and outperformed traditional methods in certain tasks

- Knowledge Graph Enriched BERT(KGEB):
  - Integrates external knowledge from a local knowledge graph, improving the performance of text classification tasks.

  - Document or Topic Classification:
    - Involves categorizing financial documents or texts into predefined topics or themes.

    - The FinBERT has been used for extracting and categorizing topics from financial reports and disclosures.

    - ESG information classification involves identifying and classifying ESG-related data from multiple sources.

    - Recent advancements in LLMs and knowledge graph integration have significantly improved the accuracy and efficiency of these classification tasks.

- Sentiment Analysis:
  - This is a vital aspect of Natural Language Processing and holds significant importance in financial applications, where understanding market sentiment can lead to impactful forecasting and actions.
- Pre-LLM Sentiment Analysis:
  - Lexicon-Based Methods:
    - The early approaches relied on predefined word lists to infer sentiment.

- These methods are simple and interpretable but struggle with context-dependent sentiments like sarcasm.
- This is applied effectively in analyzing investor sentiment from financial news or social media.

- Machine Learning Methods
  - Financial Sentiment Analysis advanced with machine learning using supervised and unsupervised techniques.
  - These methods capture complex patterns but require large datasets for training.

- Embedding-Based Methods:
  - Marked by the introduction of word embeddings such as Word2Vec capturing contextual and semantic word relationships

- Significantly improve the accuracy of sentiment analysis but depend on large datasets and can perpetuate biases.

- Sentiment Analysis with LLMs:
  - Modern models like ChatGPT and BERT excel in understanding complex financial language are are adept at handling informal expressions like emojis and specialized terminology.
  - They can process images, audio, and video (multimodal data) allowing for comprehensive sentiment analysis in financial contexts.

- Social Media and News:
  - Social media platforms and financial blogs have real-time unstructured data reflecting public sentiment on financial markets and stocks.
  - News outlets offer more objective content, originating from the established media.
  - Advanced LLMs (ChatGPT, BERT) outperform earlier models (GPT-2) in sentiment analysis of news providing more accurate stock market predictions and trading strategies.

- Corporate Disclosures:
  - Earnings Calls - Insights into financial health, strategic direction, and management's perspective. Sentiment analysis of transcripts can influence investor decisions and market perceptions.
  - Corporate Communications - Official statements, press releases, and announcements influence stakeholder perceptions. LLMs streamline and

clarify disclosures, revealing true insights by reducing redundant information.
  ○ Regulatory Filings and Legal Documents - Essential for compliance and transparency. LLMs identify sentiment-related information in complex documents, predicting stock movements and influencing financial markets.

● Market Research Reports:
  ○ These reports encompass data on economic indicators, industry analysis, and consumer behavior.
  ○ Analyst Reports and Investment Research: Provide detailed analysis and recommendations on securities, helping investors understand market trends and potential investment opportunities.
  ○ Analyst Ratings: Concise evaluations of a security's future performance, guiding investors with "buy," "hold," or "sell" recommendations based on rigorous financial analysis.

● Policy and Economic Indicators:
  ○ FOMC Meeting Minutes: Key source for understanding the U.S. Federal Reserve's monetary policy stance, offering insights into economic outlook, inflation expectations, and potential interest rate changes.
  ○ ECB Policy Decisions: Impact financial markets significantly through interest rate changes and asset purchase programs, closely monitored by investors and analysts.
  ○ Other Economic Indicators: Non-farm payroll data, unemployment rates, inflation rates, and GDP growth provide insights into market dynamics. Applying LLMs to these indicators offers advanced sentiment analysis and impacts financial market predictions.