Large Language Models for Financial and Investment Management: Applications and Benchmarks

Yaxuan Kong, Yuqi Nie, Xiaowen Dong, John M. Mulvey, H. Vincent Poor, Qingsong Wen, and Stefan Zohren

Yaxuan Kong
is a PhD student in the Department of Engineering Science at the University of Oxford in Oxford, UK.
yaxuan.kong@eng.ox.ac.uk

Yuqi Nie
is a PhD student in the Department of Electrical and Computer Engineering at Princeton University in Princeton, NJ.
ynie@princeton.edu

Xiaowen Dong
is an associate professor in the Department of Engineering Science at the University of Oxford in Oxford, UK.
xiaowen.dong@eng.ox.ac.uk

John M. Mulvey
is a professor in the Department of Operations Research and Financial Engineering at Princeton University in Princeton, NJ.
mulvey@princeton.edu

H. Vincent Poor
is the Michael Henry Strater University Professor in the Department of Electrical and Computer Engineering at Princeton University in Princeton, NJ.
poor@princeton.edu

Qingsong Wen
is the head of AI research and chief scientist at Squirrel Ai Learning in Bellevue, WA.
qingsongedu@gmail.com

Stefan Zohren
is an associate professor in the Department of Engineering Science at the University of Oxford in Oxford, UK.
stefan.zohren@eng.ox.ac.uk

KEY FINDINGS

- Thousands of studies are being conducted to evaluate the usefulness of LLMs in finance. Many of these efforts are at an early stage, so consensus has not yet formed regarding which applications are the most likely to impact current practices.
- LLMs can tailor financial advice to specific individuals and organizations to improve their ability to achieve investment goals. Similarly, investment information, such as fund historical performance, is readily available through simple queries.
- Investment performance can be enhanced by speedily deploying sentiment analysis.
- LLMs can carry out tasks such as constructing programs for asset allocation and asset–liability management.

ABSTRACT

The rapid evolution and unprecedented advancements in large language models (LLMs) have ushered in a new era of innovation in the realm of machine learning, with far-reaching implications for the finance and investment management sectors. These models have exhibited remarkable prowess in contextual understanding, processing vast and complex datasets, and generating content that aligns closely with human preferences. The transformative potential of LLMs in finance has catalyzed a surge of research and applications. As the integration of LLMs into financial practices continues to accelerate, there is an urgent need for a systematic examination of their diverse applications, methodologies, and impact, which necessitates a comprehensive review and synthesis of recent developments in this rapidly evolving field. This article aims to bridge the gap between cutting-edge artificial intelligence technology and its practical implementation in finance, providing a robust framework for understanding and leveraging LLMs in financial contexts. The authors explore the application of LLMs on various financial tasks, focusing on their potential to transform traditional practices and drive innovation. The article is highlighted for categorizing the existing literature into key application areas, including linguistic tasks, sentiment analysis, financial time series, financial reasoning, and agent-based modeling. For each application area, the authors delve into specific methodologies, such as textual analysis, knowledge-based analysis, forecasting, data augmentation, planning, decision support, and simulations. Furthermore, the article provides a comprehensive collection of datasets, benchmarks, and useful code associated with mainstream applications, offering valuable resources for researchers and practitioners. The authors hope their work can help facilitate the adoption and further development of LLMs in finance and investment management.

The domains of finance and investment management have always been characterized by complexity, uncertainty, and rapid evolution. With the advent of technology, the integration of advanced computational models in finance has gained significant momentum (Mulvey et al. 2022). Among these advancements, large language models (LLMs) have emerged as a powerful tool, demonstrating remarkable capabilities in understanding context, processing vast amounts of data, and generating human-like text. The application of LLMs in finance promises to transform traditional practices, drive innovation, and unlock novel opportunities across various financial tasks.

LLMs, such as generative pretrained transformer (GPT) series, BERT, and finance-specific variants like FinBERT, have shown impressive performance in natural language processing (NLP) tasks. These models leverage sophisticated algorithms and extensive pretraining on vast datasets to achieve advanced contextual understanding, customization capabilities, and scalability for real-time analysis. Their ability to detect complex emotional states and provide accurate interpretations makes them particularly valuable in the financial sector, where understanding market sentiment and making informed decisions are crucial.

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. For instance, in *linguistic tasks*, LLMs excel in summarizing and extracting key information from extensive financial documents, thereby streamlining complex financial narratives into concise summaries and enabling more efficient information processing. *Sentiment analysis*, one of the most crucial applications in finance, has been widely explored for decades. The advances of LLMs have made them pivotal in quantifying market sentiment from financial news, social media, and corporate disclosures, thereby providing critical insights that influence market movements and investment decisions. Additionally, LLMs have shown potential capabilities in *financial time series analysis*, including forecasting market trends, detecting anomalies, and classifying financial data, although their efficacy remains under debate. These models aim to enhance prediction accuracy and robustness by leveraging their deep learning architecture to capture complex temporal dependencies and patterns within financial datasets. One of the most promising areas of research where LLMs distinctly surpass previous deep learning methods is their capability of reasoning, which enables them not only to fit the data but also to emulate reasoning processes similar to human cognition. In *financial reasoning*, LLMs support financial planning, generate investment recommendations, and assist in decision-making by processing and synthesizing vast amounts of financial data from diverse sources. Leveraging their ability to imitate human decision-making processes, LLMs are further applied in *agent-based modeling*. This application extends the reasoning capabilities of LLMs to interactions among agents and their environments, markets, and humans, enabling the simulation of market behaviors, economic activities, and the dynamics of financial ecosystems.

Recently, several surveys have explored the applications of LLMs in the financial domain. For instance, Lee et al. (2024) present an overview of financial LLMs from the model perspective. Li, Wang, Ding, and Chen (2023) review the current approaches employing LLMs in finance and propose a decision framework to guide their adoption. Dong, Stratopoulos, and Wang (2024) provide a scoping review on ChatGPT and related LLMs in the fields of accounting and finance. Zhao, Liu et al. (2024) focus on the integration of LLMs into a variety of financial tasks. Despite these contributions, existing surveys often lack a deep dive into the practical challenges and opportunities specific to finance, or they focus primarily on the technical aspects without addressing the broader implications for financial decision-making and industry practices.

This survey aims to fill these gaps by not only reviewing the state of the art but also presenting a detailed analysis of innovative applications and useful benchmarks. Our work uniquely positions itself by providing a holistic view that is driven by *real-world applications in finance*, thus offering valuable insights for both researchers and practitioners.

LINGUISTIC TASKS: TEXTUAL WORK

Many earlier models, such as those based on recurrent neural networks (RNNs), specifically long short-term memory (LSTM), have demonstrated a capacity for achieving a degree of language understanding over text sequences and performing textual work (Lipton, Berkowitz, and Elkan 2015). Due to these models' architectural constraints, however, they struggled with long-term dependencies. Specifically, they encountered challenges in maintaining context over long text sequences, understanding complex expressions, dealing with large datasets, and handling unstructured data efficiently (Staudemeyer and Morris 2019; Kong et al. 2024). This limitation is particularly evident when applying in the financial sector, where the volume of documentation is vast, and the need for accurate and concise summaries is critical (Zmandar et al. 2021).

LLMs, which leverage the transformer model architecture, conversely, have significantly advanced the field's capabilities (see Exhibit 1 for LLM applications on linguistic tasks). The transformer architecture, characterized by its innovative self-attention mechanism, allows LLMs to process, understand, and generate text based on massive datasets on which they have been trained (Hadi et al. 2024; Raiaan et al. 2024). This breakthrough is instrumental in overcoming the challenges faced by earlier models. By efficiently managing long-term dependencies and contextual information over large volumes of text, LLMs can streamline complex financial narratives into concise summaries and extract relevant information (Hadi et al. 2024; Raiaan et al. 2024). This process retains essential insights and enables more efficient information processing.

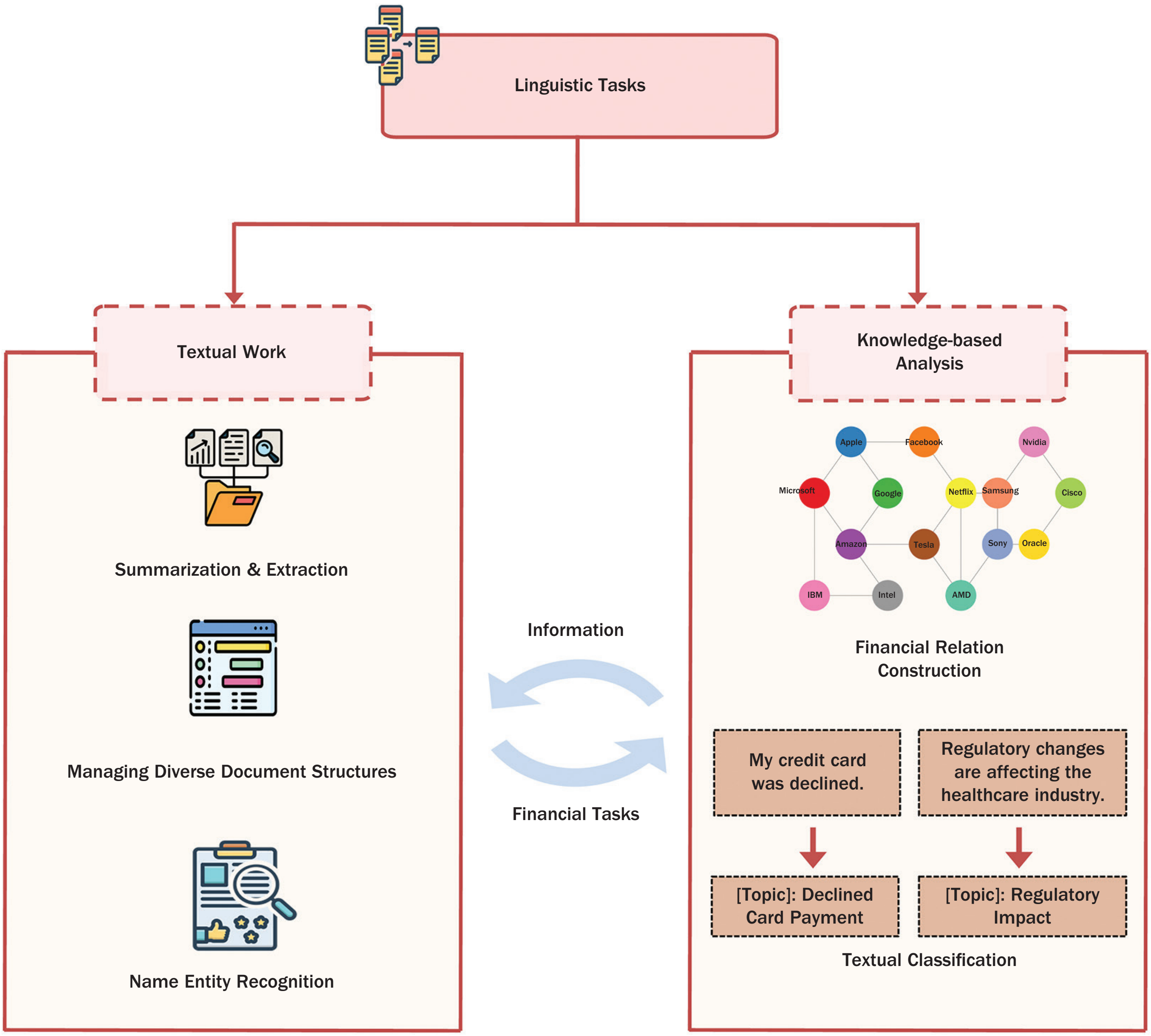
Summarization and Extraction

Recent research has effectively utilized LLMs to summarize and extract financial document information (Abdaljalil and Bouamor 2021; La Quatra and Cagliero 2020; Ni, Li, and Li 2023). Given that these financial documents are often lengthy, which can exceed the token limits of many LLMs, various studies have introduced frameworks by dividing long documents into shorter segments or utilized specific models to address the challenges of processing extensive financial texts (Xia et al. 2022; Vanetik, Podkaminer, and Litvak 2023). Recently, Yepes et al. (2024) propose an expanded approach to chunking documents for retrieval augmented generation (RAG) by using structural elements rather than just paragraphs, which improves chunk size determination without tuning. Furthermore, some papers propose segmenting long reports into 10 distinct sections, such as management's discussion and analysis, financial highlights, and business overview, to streamline the summarization process (Shukla et al. 2022; Shukla et al. 2023). Similarly, Khanna et al. (2022) utilize the Longformer-Encoder-Decoder (LED) model, a transformer architecture first introduced by Beltagy, Peters, and Cohan (2020), which employs a self-attention mechanism scalable with sequence length, making it suitable for analyzing long financial reports.

Beyond handling the long size of the document, research has expanded into multilingual and domain-specific challenges. This includes summarizing financial documents across multiple languages (Foroutan et al. 2022); customizing language

EXHIBIT 1

Illustration of Various Linguistic Tasks in Finance



models to tackle the adaptation challenges to Japanese financial terminology (Suzuki et al. 2023); automating the finetuning process for text summarization models in the cryptocurrency domain without requiring human annotation (Avramelou et al. 2023); adopting multitask learning strategies to classify, detect, and summarize financial events (Li and Zhang 2021); addressing the challenge of ensuring accuracy and reducing errors in financial information extraction (Sarmah et al. 2023); extracting information from annual reports to enhance stock investment strategies (Gupta 2023).

Managing Diverse Document Structures

Despite the effectiveness of LLMs in handling textual financial data, they often face challenges with PDF document formats that incorporate images, charts, and tables.

This challenge may arise from their primarily text-based nature, which finds it difficult to interpret the complex spatial layouts crucial for understanding such multimodal documents (Li, Gao, Wu, and Vasarhelyi 2023). One simple approach to this issue involves converting PDF files into machine-readable plain text. For instance, in the automated financial information extraction (AFIE) framework proposed by Yue et al. (2023), tables are transformed into text using PLAIN serialization. This method uses spaces and newline characters to separate cells and rows, respectively. This effectively integrated table data with regular paragraphs for LLMs to process uniformly.

However, this conversion process may change the document's spatial layout and potentially lead to the loss of crucial information embedded in charts or tables. To address this, the team at JP Morgan developed DocLLM (Wang, Raman et al. 2023), a layout-aware generative language model specifically designed for multimodal document understanding. DocLLM utilizes bounding box information to understand the spatial arrangement of elements within the documents. It enhances document understanding by modifying the attention mechanism in transformers to concentrate on the cross-alignment between textual and spatial modalities.

Name-Entity Recognition

Name-entity recognition (NER) is a subtask of information extraction and plays a crucial role in extracting meaningful information from various financial sources (Li et al. 2020; Ehrmann et al. 2023). In the financial domain, it is used to extract specific entities such as company names, financial terminology, stock symbols, financial indicators, and monetary values from news articles, financial reports, and market summaries (Swaileh et al. 2020). This information is crucial for financial downstream tasks, such as industry classification, sentiment analysis, credit scoring, fraud detection, and regulatory compliance reporting (Alvarado, Verspoor, and Baldwin 2015).

Traditionally, NER is approached through rule-based methods, machine learning techniques, or deep learning techniques (Nadeau and Sekine 2007). Rule-based methods depend on handcrafted linguistic and grammatical rules. They offer high precision for well-defined patterns but suffer from limited scalability (Li et al. 2020). Machine learning techniques include both supervised and unsupervised approaches. Supervised approaches utilize a comprehensive set of engineered features, such as word-level characteristics and list lookups, alongside machine learning algorithms, such as hidden Markov models (Eddy 1996), decision trees (Quinlan 1986), and support vector machines (Hearst et al. 1998), to identify and classify entities in text. Unsupervised learning approaches extract and classify named entities by employing clustering, leveraging lexical resources and patterns, and analyzing corpus statistics (Nadeau and Sekine 2007). While machine learning offers flexibility and can handle diverse data types, it relies heavily on the availability of labeled data for supervised learning and can lack interpretability in unsupervised learning (Li et al. 2020). Deep learning methods utilize advanced architectures such as bidirectional long short-term memory (BiLSTM) networks, self-attention-based transformers, and conditional random fields (CRF) for tag decoding to effectively learn and represent both word- and character-level features from large datasets. These approaches significantly enhance model performance by enabling the capture of complex patterns and long-range dependencies in text (Li et al. 2020).

With the emergence of deep learning methods, LLMs are now increasingly used in NER within the financial domain (Pakhale 2023; Wang, Pan et al. 2023). The ability of LLMs to leverage vast pretrained knowledge and sophisticated language understanding can significantly enhance the accuracy and efficiency of entity recognition in complex financial texts (Pakhale 2023). Recently, Hillebrand et al. (2022) propose KPI-BERT, a new system that utilizes advanced techniques NER and relation extraction

(RE) to identify and connect key performance indicators (KPIs), such as "revenue" or "interest expenses," within German financial documents. This system relies on an end-to-end trainable architecture based on BERT. It combines an RNN with conditional label masking for sequential entity tagging, followed by relation classification. Further research has utilized LLMs for NER to improve the efficiency and accuracy of XBRL (eXtansive Business Reporting Language) tagging (Loukas et al. 2022); identify similar peer companies (Covas 2023); detect key entities of negative news information (Zhao, Zheng, and Zhang 2021); and extract relevant phrases for entities (Gupta et al. 2021).

Despite their demonstrated exceptional generalization capabilities, LLMs sometimes come with high training and inference costs, especially when processing long documents. To address these, Zhou et al. (2023) propose UniversalNER, a model that employs targeted distillation with mission-focused instruction tuning to train cost-efficient student models for open NER. This approach not only reduces the computational burden but also achieves remarkable NER accuracy without direct supervision.

LINGUISTIC TASKS: KNOWLEDGE-BASED ANALYSIS

In financial text analysis, summarizing and extracting key information from documents is crucial for quickly understanding and processing important data within lengthy and complex texts (Xue et al. 2023). Following the extraction of pertinent information, the next step involves utilizing this information for solving downstream financial tasks. This section will introduce two main activities central to this application: constructing financial relationships and textual classification. These efforts are vital for leveraging the extracted information to enhance decision-making and analytical processes in the finance sector.

Financial Relation Construction

Constructing financial relationships, particularly through the use of knowledge graphs, represents a powerful methodology for organizing and making sense of the extracted entities and their interrelations from extensive and complex financial datasets (van Zwam et al. 2020). Knowledge graphs consist of interconnected descriptive structures about entities (objects, events, people, etc.), the attributes of those entities, and the relationships that link them together. This framework offers a structured way of representing relationships within data and enables sophisticated analyses to be derived from them (Jiang et al. 2023; Pan et al. 2023).

Upon the identification and extraction of entities (such as companies, individuals, financial instruments, events, etc.), along with the relationships among these entities (such as ownership, transactions, legal disputes, etc.), this information can be systematically organized into a graph format for further construction. Within a knowledge graph, entities are represented as nodes, and relationships are denoted as edges that connect these nodes. This structure provides a visual and programmable method to explore and comprehend the connections among different entities within the financial ecosystem. With the construction of the knowledge graph, financial analysts and systems can employ graph analytics and machine learning algorithms to discover insights, recognize patterns, and predict future events (Pan, Luo et al. 2024).

Recent advancements in LLMs have led researchers to explore the potential of using information extracted by LLMs to construct and analyze knowledge graphs in the financial sector (Trajanoska, Stojanov, and Trajanov 2023; Ouyang et al. 2024; Wang, Sun, Chen, and Cui 2022). Notably, Trajanoska, Stojanov, and Trajanov (2023) generate a knowledge graph by leveraging LLMs to extract structured environmental, social, and governance (ESG) information from sustainability reports, using a format of triples consisting of node-edge-node, to enable deeper analysis and understanding of corporate sustainability practices. Similarly, Cheng et al. (2022) develop a semantic-entity interaction module. This module combines a language model with a conditional random field (CRF) layer to comprehend the interaction between entities and their semantic contexts in texts. It automatically constructs financial knowledge graphs from brokerage research reports without the need for explicit financial knowledge or extensive manual rules.

Moreover, financial research analysts often face challenges in identifying critical documents, key entities, and important events during their research on complex financial subjects. Mackie and Dalton (2022) tackle these issues by developing automated methods to create detailed, query-specific knowledge graphs from documents and entities.

As illustrated above, knowledge graphs have demonstrated their utility in information retrieval. A special case within this domain is the translation of natural language (NL) into graph query language (GQL). This process enhanced querying experiences by leveraging the relational data within knowledge graphs, offering advantages over traditional text-to-SQL methods. However, this approach is challenged by the complexity of accurately mapping NL to GQL syntax and the lack of domain-specific examples, making it difficult to fine-tune LLMs for precise alignment with graph databases in specialized fields (Pan et al. 2023). To address this, Liang, Tan et al. (2024) develop a pipeline that employs LLMs to generate NL-GQL pairs from financial graph databases without labeled data. This process involved creating template pairs with ChatGPT and refining them through a self-instruction method. Subsequently, LLMs were fine-tuned with these pairs using the low-rank adaptation of LLMs (LoRA) technique to align the models with the specific knowledge contained in graph databases.

Knowledge graphs can also be used to significantly enhance question-answering systems. Wang, Lipka et al. (2024) introduce an innovative knowledge graph prompting (KGP) for multidocument question answering (MD-QA). Their approach constructs a knowledge graph from multiple documents, highlighting semantic or lexical relationships among passages or document structures. An LLM-based graph traversal agent then uses this knowledge graph to gather contextually relevant information, thereby enhancing the LLM's accuracy in answering questions.

Another beneficial aspect of knowledge graphs is their ability to be enriched over time through the use of LLMs. Li (2023) presents FinDKG, a dynamic knowledge graph with LLMs used in financial domain. FinDKG incorporates a temporal layer in its structure, which allows it to reflect and adapt to changes in financial markets, economic indicators, and thematic trends. This dynamic approach provides valuable insights for thematic investing, making it possible to identify and leverage long-term industry shifts and economic trends for strategic investment decision-making.

There exist other financial relation extraction studies using LLMs, though not necessarily for knowledge graph construction (Ok 2023; Kaur et al. 2023; Chai et al. 2023; Tian, Zhao, and Ren 2019). Ghosh et al. (2023) propose the Mask One At a Time (MOAT) framework, which masks one entity at a time, extracts contextual embeddings using a domain-specific language model (SEC-BERT), and combines these embeddings with additional features to train a neural network for accurately classifying relationships among financial entities. Similarly, Rajpoot and Parikh (2023) employ in-context learning with GPT models, utilizing both a learning-free dense retriever (KNN with OpenAI embeddings) that relies on the similarity of embeddings to find the most relevant examples and a learning-based retriever trained to select the most similar example in the training set for each test example by estimating the probability of the output given the input and a candidate training example as the prompt. Focusing on multitype Chinese financial event relation extraction, Wan et al. (2023) propose the CFERE framework, which employs a core verb chain for event identification, constructs a syntactic semantic dependency parsing graph to combine events into pairs, and enhances BERT with an event core embeddings layer to capture semantic meanings. These studies demonstrate the potential of LLMs and innovative approaches in advancing financial relation extraction, ultimately contributing to the research value of making use of financial information and helping investors make better investment decisions.

Textual Classification

Textual classification plays a crucial role in organizing and understanding large volumes of unstructured data within the financial domain. This classification task can be further categorized into several subtasks, such as industry/company classification and document/topic classification. By effectively classifying and organizing this information, businesses and researchers can extract valuable insights and make informed decisions. The utilization of these classification techniques, in conjunction with the establishment of financial relationships, is essential for leveraging the extracted information to enhance decision-making and analytical processes within the finance sector.

Company or industry classification involves grouping companies into distinct categories based on shared characteristics, such as business activities and market performance, with the aim of creating coherent and differentiated groups. Identifying similar company profiles is a fundamental task in finance, with applications spanning investment portfolio construction, securities pricing, and financial risk attribution. Traditionally, financial analysts have relied on industry classification systems, such as the Global Industry Classification System (GICS), the Standard Industrial Classification (SIC), the North American Industry Classification System (NAICS), and the Fama–French (FF) model, to identify companies with similar profiles (Rizinski et al. 2024). However, these systems do not provide a means to rank companies based on their degree of similarity and require time-consuming, effort-intensive manual analysis and data processing by domain experts (Rizinski et al. 2024).

Recently, a team at BlackRock (Vamvourellis et al. 2023) explored a novel approach to company classification using LLMs. They investigated the use of pretrained and fine-tuned LLMs to generate company embeddings based on business descriptions from SEC filings. Their study aimed to assess the embeddings' ability to reproduce GICS classifications, benchmark LLM performance on various downstream financial tasks, and examine the impact of such factors as pretraining objective, finetuning, and model size on embedding quality. The results showed that LLM-generated embeddings, particularly those from fine-tuned Sentence-BERT models, could accurately reproduce GICS sector and industry classifications and outperform them on tasks such as identifying similar companies based on return correlations and explaining cross-sectional equity returns.

Interestingly, knowledge graphs can also be used to enrich industry classification and improve the performance of domain-specific text classification tasks. Wang et al. (2021) propose a novel Knowledge Graph Enriched BERT (KGEB) model that integrates external knowledge from a local knowledge graph with word representations. They demonstrated the effectiveness of their approach by constructing a large dataset based on companies listed on the Chinese National Equities Exchange and Quotations (NEEQ) and showing that the KGEB model outperforms competitive baselines, including graph convolutional network, logistic regression, TextCNN, BERT, and K-BERT, achieving an accuracy of 91.98% and an F1 score of 90.89%.

Document or topic classification is another crucial subtask within the broader scope of textual classification in the financial domain. This task involves categorizing financial documents or texts, such as news articles (Mishra 2023; Nugroho, Sukmadewa, and Yudistira 2021) or company filings (Arslan et al. 2021; Loukas et al. 2023), into predefined topics or themes. Alias et al. (2023) propose a novel approach that utilizes the FinBERT model to extract and categorize relevant topics of key audit matters (KAM) from the annual reports of publicly listed companies in Bursa Malaysia. Similarly, Burke et al. (2023) fine-tune the FinBERT model to classify accounting topics within three unlabeled financial disclosures, including custom notes to the financial statements, the Management's Discussion and Analysis section, and the risk factor section.

Another important classification task in the financial domain involves categorizing ESG information. This task requires identifying and classifying ESG-related data, such as carbon emissions, diversity and inclusion, and corporate governance practices, from multiple sources including corporate sustainability reports, news articles, and social media posts. In a recent study, Lee and Kim (2023) propose an ESG classifier that can discriminate ESG information by fine-tuning a pretrained language model. The classifier was trained on a manually labeled dataset constructed from sustainability reports of Korean companies across five sectors and achieved a classification accuracy of 86.66% for the four-class classification problem (environment, social, governance, and neutral). Similarly, Mehra, Louka, and Zhang (2022) develop a domain-specific language model called ESG-BERT to enhance the classification of ESG-related text by fine-tuning BERT's pretrained weights using ESG-specific text and further fine-tuning the model for classification tasks.

Textual classification techniques, including industry/company classification and document/topic classification, play a vital role in organizing and understanding large volumes of unstructured data within the financial domain. Recent advancements in LLMs and knowledge graph integration have significantly improved the accuracy and efficiency of these classification tasks. The successful application of these techniques can further provide valuable insights and support informed decision-making in various financial contexts, such as investment portfolio construction, risk assessment, and ESG analysis.

SENTIMENT ANALYSIS

Sentiment analysis emerges as a crucial component within the domain of NLP and is one of the most important tasks in financial applications. It involves the quantitative exploration of opinions, sentiments, subjectivity, and emotions articulated in textual data (Tan, Lee, and Lim 2023; Bordoloi and Biswas 2023; Fabozzi 2024). This task acquires particular significance within financial applications, where the interpretation of market sentiment can lead to impactful forecasting and actions (Mishev et al. 2020). Its evolution mirrors the broader advancements in NLP, transitioning from rule-based systems to sophisticated machine learning models, and more recently, to deep learning approaches that leverage large pretrained language models.

Pre-LLM Sentiment Analysis

First, we outline the significant milestones in sentiment analysis in this section, leading up to the era before LLMs such as ChatGPT and BERT revolutionized the field. Additionally, it highlights key applications within the financial domain, demonstrating the impact of sentiment analysis on various applications.

Lexicon-based methods. Early sentiment analysis relied on lexicon-based approaches, where the sentiment of a text was inferred based on the presence of predefined words associated with positive or negative sentiments. These methods, simple yet effective for certain applications, include the general inquirer (Stone, Dunphy, and Smith 1966) and linguistic inquiry and word count (LIWC) lexicons (Pennebaker, Francis, and Booth 2001), SO-CAL (Taboada et al. 2011), and Loughran and McDonald's (LM) word lists (Loughran and McDonald 2011). One of the strengths of lexicon-based methods is their simplicity and interpretability. However, their performance can be limited by the context-dependency of sentiment expressions and the inability to capture the sentiment expressed by complex linguistic constructs, such as sarcasm or irony. Despite these limitations, lexicon-based methods have been effectively applied in finance, particularly in analyzing investor sentiment from financial news or social media content (Sohangir, Petty, and Wang 2018; Yekrangi and Abdolvand 2021; Consoli, Barbaglia, and Manzan 2022).

Machine learning methods. With the advent of machine learning (ML), financial sentiment analysis (FSA) experienced significant advancements. ML-based methods can be broadly categorized into supervised and unsupervised learning. When doing FSA, supervised learning approaches require labeled data and include such techniques as support vector machines (SVM) (Chiong et al. 2018), naive Bayes (Kalra and Prasad 2019), KNN (K-Nearest Neighbor; Kirange and Deshmukh 2016), random forests (Dickinson and Hu 2015), and multilayer perceptrons (MLPs; Valencia, Gómez-Espinosa, and Valdés-Aguirre 2019). Unsupervised learning, in contrast, does not require labeled data and typically involves clustering techniques to discern sentiment (Yadav et al. 2020). In finance, ML has been used to predict market movements based on sentiment in financial news and social media, showcasing its ability to capture financial sentiment nuances (Renault 2020). Machine learning methods offer the advantage of being able to capture complex patterns in data that are not apparent to lexicon-based methods. However, they require large datasets for training, and the versatility is limited on a specific domain.

Embedding-based methods. The introduction of word embeddings marked a significant milestone in general sentiment analysis. Embedding-based methods represent textual information in a high-dimensional space where semantically similar words are closer together. This representation captures not only the sentiment but also the context of words, leading to improved performance in sentiment analysis tasks. Mikolov et al.'s (2013) introduction of Word2Vec in 2013 was a pioneering development in this domain. Word2Vec employs neural networks to learn word associations from large datasets, generating embeddings that capture a wide array of linguistic relationships and nuances. The innovative aspect of Word2Vec lies in its ability to learn high-quality word vectors from vast datasets efficiently. It offers two architectures for this purpose: Continuous Bag of Words (CBOW) and Skip-gram. CBOW predicts target words from context words, while Skip-gram does the opposite, predicting context words from a target word, making it particularly effective for capturing semantic and syntactic word relationships.

Subsequent to Word2Vec, several other embedding models have emerged, further advancing the field. Notable among these are Global Vectors for Word Representation (GloVe; Pennington, Socher, and Manning 2014), which introduces an unsupervised learning algorithm for obtaining vector representations of words through aggregating global word–word co-occurrence statistics from a corpus; FastText (Bojanowski et al. 2017), which extends Word2Vec to consider subword information, thereby enhancing the representation of rare words; and Embeddings from Language Models (ELMo; Sarzynska-Wawer et al. 2021), which leverages bidirectional language models to generate contextually enriched word embeddings.

Beyond word-level embeddings, there has been a push toward capturing longer contextual dependencies. An exemplar in this area is Doc2Vec, also known as Paragraph Vector, introduced by Le and Mikolov (2014). Doc2Vec extends the Word2Vec paradigm to support document-level embeddings, enabling the capture of document-wide contextual information that is crucial for tasks requiring comprehension of extended textual content. By learning fixed-length feature representations from variable-length pieces of texts, Doc2Vec facilitates a deeper understanding of document semantics, thereby broadening the applicability of embedding techniques in sentiment analysis and beyond.

Embedding-based methods have the advantage of capturing contextual complexity and semantic relationships among words, significantly improving the accuracy of sentiment analysis. This has made them popular in FSA as well. Sohangir et al. (2018) highlight the effectiveness of these methods in financial domains, demonstrating their ability to extract sentiment from large volumes of unstructured financial data with high accuracy.

However, they are not without drawbacks. A notable limitation is their dependency on large datasets for training, which might not always be feasible in specialized domains. Additionally, while adept at semantic understanding, they may overlook slight differences in syntax and require retraining to adapt to new language uses or vocabularies. Pretrained embeddings can also perpetuate biases present in their training data, leading to potential issues in fairness and representation. Despite these challenges, embedding-based methods are crucial in advancing natural language understanding and have paved the way for large language models like BERT and GPT-3, which build on these embeddings to achieve state-of-the-art NLP performance.

Sentiment Analysis with LLMs

The advent of ChatGPT and other LLMs represents a pivotal milestone in the domain of FSA. Nowadays, these models have demonstrated their effectiveness in numerous tasks and offer several unique advantages for FSA applications.

First, LLMs excel in deciphering the complexities of financial language, adeptly navigating informal expressions, emojis, memes, and specialized terminology across social media and financial blogs (Deng et al. 2023; Steinert and Altmann 2023; Chen and Xing 2023; Vamossy and Skog 2023; Jeong 2024; Mumtaz and Mumtaz 2023). Their proficiency in identifying subtleties like irony, sarcasm, and sector-specific jargon is vital for accurately analyzing sentiments across various formats, from social media posts to comprehensive financial reports (Băroiu and Trăuşan-Matu 2023; Wu et al. 2023).

Second, LLMs' ability and great potential to process multimodal data, including images, audio, and video, are essential for comprehensive sentiment analysis in financial contexts like earnings calls (Cook et al. 2023) and FOMC meetings (Curti and Kazinnik 2023). This capability allows for the integration of nonverbal cues and visual data into the sentiment analysis process (Bhatia et al. 2024).

Third, LLMs' ability to process extensive documents enables thorough analysis of detailed financial reports and lengthy articles, ensuring no sentiment-bearing information is overlooked. This feature is particularly beneficial for evaluating the sentiments expressed in annual reports, earnings transcripts, and extensive financial narratives (Kim, Muhn, and Nikolaev 2023).

Moreover, LLMs exhibit enhanced resilience to adversarial attacks or deceptive information tactics that could be encountered in FSA tasks. Their advanced algorithms and broader contextual understanding help in identifying and mitigating misleading or manipulative sentiment indicators, enhancing the reliability of sentiment analysis outcomes. Leippold (2023) highlights the contrast between traditional keyword-based sentiment analysis methods and LLMs in the face of adversarial attacks. The research involved using GPT-3 to substitute negative words with synonyms to assess model robustness, showcasing FinBERT's enhanced resilience against adversarial attacks over traditional keyword-based methods.

Data-Driven Applications of LLMs in FSA

We further delve into the recent advancements in the integration of LLMs within FSA, categorically analyzing the impact and contributions according to diverse data sources. We embark on this exploration by categorizing the data into four key segments:

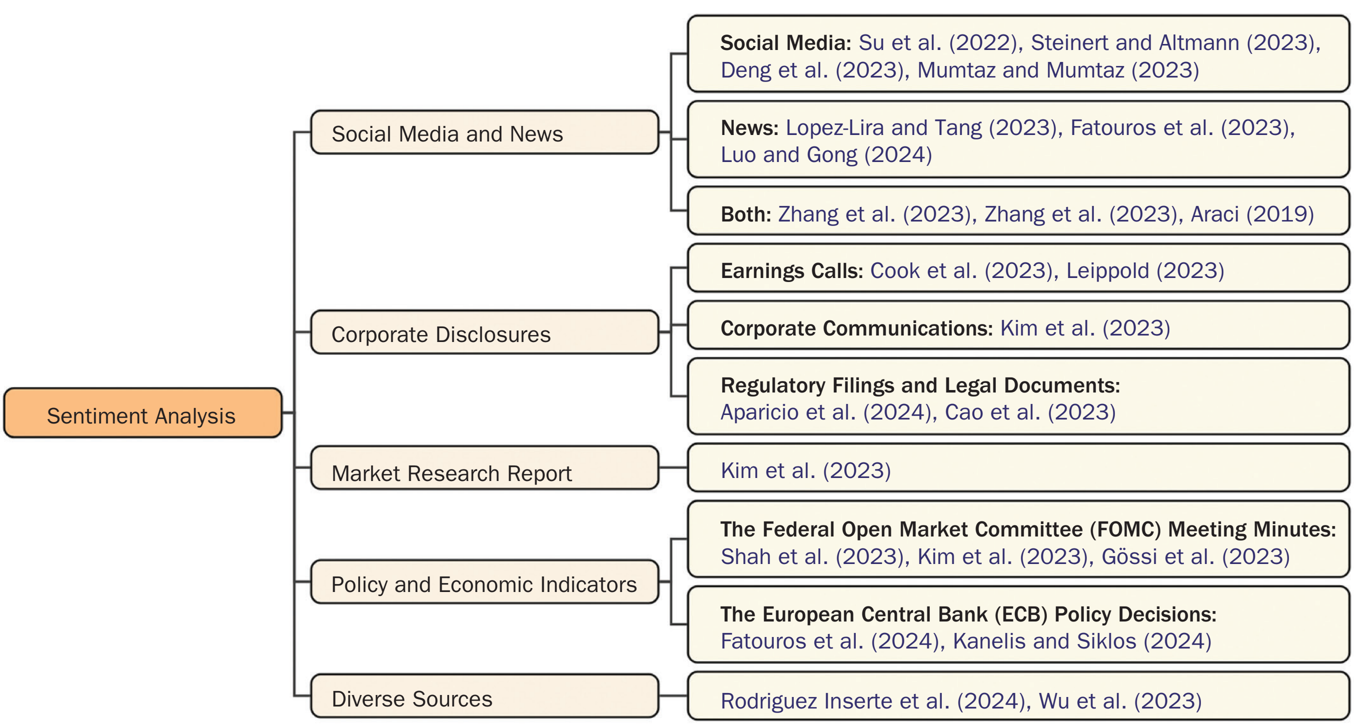
1. social media and news,
2. corporate disclosures,
3. market research reports, and
4. policy and economic indicators.

This structured approach enables a comprehensive understanding of how LLMs have revolutionized the domain of FSA, offering unprecedented insights and analytics capabilities. The collection of representative articles for sentiment analysis tasks is provided in Exhibit 2.

Social media and news. Social media platforms such as X (formerly named Twitter), general online forums like Reddit, and finance-specific forums such as StockTwits, along with financial blogs and microblogs, have become rich sources of data for FSA. These platforms are crucial due to their rich repositories of real-time, unstructured textual content that mirrors public sentiment regarding financial markets, specific stocks, and the overall economic environment. The immediacy and public nature of the discussions on these platforms make them an invaluable resource for capturing the mood of the market, which can be predictive of future market movements. Su, Mulvey, and Poor (2022) leverage BERT for extracting sentiment and semantic insights from Twitter, facilitating improved covariance estimation and enhancing portfolio optimization. The integration of text-derived covariance data into mean–variance optimization

EXHIBIT 2

Selected Representative Articles for Sentiment Analysis Tasks in Finance, Categorized by Various Data Sources



resulted in superior performance in this work, especially during the COVID-19 crash period. Furthermore, Steinert and Altmann (2023) employ GPT-4 for sentiment analysis of microblogging messages on the StockTwits platform, outperforming the naive buy-and-hold strategy for Apple and Tesla stocks by a significant margin, which underscores the potential of LLMs in predicting stock price movements through sentiment analysis. Despite the efficacy of LLMs in sentiment analysis, social media sources present unique challenges, including the vast volume of information, the colloquial language often used, possible selective bias, and the presence of misinformation or inaccuracies in the messages shared, which complicate the task of accurately capturing and interpreting market sentiments (Ebrahimi, Yazdavar, and Sheth 2017).

News represents another crucial data source, which shares many similarities with social media in terms of rapid dissemination and broad reach, but generally focuses more on objective events. Contrary to the often subjective and personal nature of social media, news content typically originates from more prestigious and established media outlets, including renowned newspapers such as *The New York Times*, television broadcasters like CNN and BBC, and finance-specific publications such as *The Economist*. The credibility and professionalism of journalists and writers in these outlets lend trustworthiness to the content, albeit sometimes at the cost of timeliness. Evidence increasingly supports the advantages of post-ChatGPT LLMs over earlier approaches, particularly in analyzing the sentiment of news headlines. Lopez-Lira and Tang (2023) investigate ChatGPT's effectiveness in predicting stock market returns, illustrating its capability to accurately assign sentiment scores to headlines and outperform earlier models such as GPT-2 and BERT. Additionally, Fatouros et al. (2023) reveal that GPT-3.5 offers considerable improvements over FinBERT in analyzing forex-related news headlines. Similarly, Luo and Gong (2024) report noteworthy success with the open-source Llama2-7B model (Touvron et al. 2023), achieving performance that exceeded previous BERT-based approaches and conventional methods like LSTM with ELMo. These researches underscore the significance of advanced LLMs in decision-making and quantitative trading.

In this digital age, the phenomenon of real-time news is becoming increasingly prevalent. Distributed via live broadcasts or online platforms, these news sources manage to strike a balance between accuracy and immediacy, offering timely insights into market conditions and public events that could influence financial sentiments (Arvanitis and Bassiliades 2017). Chen, Kelly, and Xiu (2022) investigate using advanced LLMs such as BERT, RoBERTa, and OPT for sentiment analysis and stock prediction. These models significantly outperform traditional methods such as Word2vec by capturing complex text information and providing a more accurate contextual understanding. They also demonstrate that LLM-based models achieve higher Sharpe ratios and better performance. Crucially, the research reveals that news information is incorporated into stock prices with a delay due to limits to arbitrage, creating opportunities for real-time trading strategies to exploit these inefficiencies. This underscores the potential of LLMs in real-time financial text mining.

Corporate disclosures. Corporate disclosures are increasingly recognized for their significance in FSA. This section delves into three primary categories of corporate disclosures: earnings calls, corporate communications, and regulatory filings and legal documents (e.g., SEC filings), each highlighted for its importance and accompanied by pertinent studies.

Earnings calls are crucial for providing insights into a company's financial health, strategic direction, and management's perspective on performance and future prospects. The sentiment analysis of earnings calls transcripts can reveal underlying tones and sentiments that may influence investor decisions and market perceptions. Cook et al. (2023) evaluate the performance of local LLMs in interpreting financial texts, particularly focusing on analyzing the tone and content of bank earnings calls during the post-COVID-19 pandemic era and the early 2023 banking stress. They show that local LLMs are effective for analyzing financial communications, demonstrating a shift in bank earnings call content toward more homogeneity and less positive sentiment during periods of increased banking stress. Leippold (2023) demonstrates the susceptibility of financial sentiment analysis to adversarial attacks using GPT-3, highlighting the need for LLMs to ensure the reliability of AI in financial text processing.

Corporate communications encompass a wide range of official statements, press releases, and announcements made by a company to its stakeholders. The sentiment embedded within these communications can significantly affect how stakeholders perceive the company's current state and future outlook. LLMs can process these communications to assess the sentiment and identify potential market-moving information. For instance, Kim, Muhn, and Nikolaev (2023) illustrate that ChatGPT can significantly streamline and clarify corporate disclosures for investors by reducing the length and amplifying the sentiment of the content, while also revealing the prevalent issue of "bloat"—excessive, redundant, or irrelevant information in financial reports—that can obscure the true insights needed for informed investment decisions.

Regulatory filings and legal documents are essential for compliance, governance, and transparency, providing a wealth of information on a company's operations, risks, and financial condition. LLMs can process these complex documents and identify sentiment-related information, such as litigation risks, accounting irregularities, and management changes. Aparicio et al. (2024) introduce BioFinBERT, a fine-tuned language model that utilizes sentiment analysis of regulatory filings and legal documents, such as 10-Q, 10-K, 6-K, and 20-F reports, along with biotech company press releases, to execute market orders and predict stock price movements in the biotechnology sector. Another recent paper (Cao et al. 2023) investigates how corporations adjust their regulatory disclosures to be more machine-readable in the age of AI, influencing both the sentiment expressed and the speed of information dissemination in financial markets.

Market research reports. Market research reports, which encompass a wide range of data including economic indicators, industry analysis, and consumer behavior, are crucial for informed decision-making in finance. The significance of analyst reports and investment research lies in their detailed analysis and recommendations on securities, offering a profound understanding of market trends and potential investment opportunities. Analyst ratings, such as "buy," "hold," or "sell" recommendations, provide another concise evaluation of a security's future performance, serving as a valuable guide for investors. These ratings are based on rigorous financial analysis and are closely monitored by investors to assess market sentiment and make strategic investment choices (Kim, Kim et al. 2023).

Policy and economic indicators. In the field of financial sentiment analysis, particularly with respect to policy and economic indicators, a significant focus has been placed on the analysis of Federal Open Market Committee (FOMC) meeting minutes, European Central Bank (ECB) policy decisions, and other key indicators such as non-farm payroll data, unemployment rates, inflation rates, and GDP growth. These sources are critically important for understanding the market dynamics and guiding investment decisions based on the sentiment derived from policy decisions and economic reports.

The FOMC meeting minutes are an important source of information for understanding the US Federal Reserve's monetary policy stance (Rosa 2013; Smales and Apergis 2017). These minutes provide a detailed account of the discussions and deliberations that take place during FOMC meetings, shedding light on the economic outlook, inflation expectations, and potential interest rate changes (Shah, Paturi, and Chava 2023). Researchers have employed LLMs to analyze the sentiment and tone of FOMC meeting minutes. Kim, Sporer, and Handschuh (2023) discuss that while FinBERT outperforms traditional techniques in predicting negative sentiment within FOMC statements, there is a need for further enhancements and exploration of alternative approaches to optimize the analysis of FOMC texts and gain more comprehensive economic insights. Gössi et al. (2023) present a fine-tuned FinBERT model with a sentiment focus method, which significantly improves the sentiment analysis accuracy of complex financial sentences in FOMC Minutes, particularly those containing conjunctions with contradicting sentiments.

The ECB is responsible for setting monetary policy for the Eurozone, and its policy decisions have a significant impact on financial markets (Klejdysz and Lumsdaine 2023). ECB policy decisions, including interest rate changes and asset purchase programs, are closely monitored by investors and analysts (Anastasiou and Katsafados 2023; Mody and Nedeljkovic 2024). Recent research has utilized LLMs to analyze the sentiment and impact of ECB policy decisions on financial markets (Fatouros et al. 2024). Using the FinBERT model, Kanelis and Siklos (2024) reveal that sentiments from monetary policy speeches explain the tone of press conference statements, while financial stability speeches offer little explanatory power, highlighting the LLM's ability to provide detailed sentiment analysis in economic communication.

In addition to FOMC meeting minutes and ECB policy decisions, several other economic indicators and research papers are relevant to FSA. Non-farm payroll data and unemployment rates provide insights into the labor market and can have a significant impact on market sentiment (Nia et al. 2022). Inflation rates and GDP growth are also closely watched indicators, as they reflect the overall health of the economy (Shapiro, Sudhof, and Wilson 2022; Biswas et al. 2020). Applying LLMs to analyze the sentiment and impact of these economic indicators on financial markets deserves further exploration for future research.

FINANCIAL TIME SERIES ANALYSIS WITH LLMs

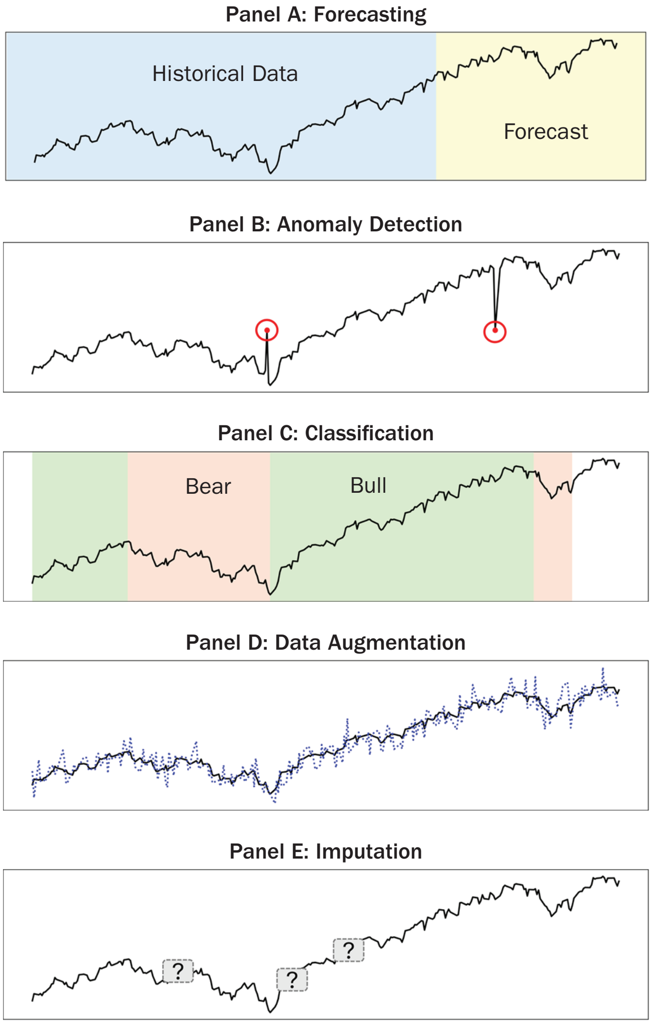
Deep learning has revolutionized time series analysis, offering powerful tools for modeling and forecasting sequential data (Lim and Zohren 2021; Wang, Nie et al. 2023). Prominent deep learning models like LSTM networks and CNNs have demonstrated significant effectiveness in capturing temporal dependencies and anomalies in time series data (Pan, Jiang, Song et al. 2024; Wang, Sun, Hu et al. 2022; Chen, Gel, and Poor 2022).

With the recent surge in popularity of LLMs, these tools are increasingly being used to assist in time series tasks, as shown in Exhibit 3 (Jiang et al. 2024; Zhang, Sun et al. 2024). They offer a multitude of auxiliary functions such as generating additional features from textual data and producing descriptive statistics, as we have discussed in the earlier section on sentiment analysis, which can enhance the accuracy of time series models by tapping into a broader spectrum of information beyond the original data.

Beyond these supportive roles, LLMs are also being employed to directly analyze time series data (Jin et al. 2024; Pan, Jiang, Garg et al. 2024), a development supported by several factors. This is primarily attributed to LLMs' ability to understand and process sequential data, which is a common trait between text and time series. Also, the prevalent transformer architecture underlying most LLMs has proven effective in various time series tasks (Zhou et al. 2022; Nie et al. 2022; Wen et al. 2022). Furthermore, LLMs exhibit remarkable multimodal capabilities, suggesting that their pretraining on vast datasets, even if solely text-based, imparts general inference and reasoning abilities beyond the specific data modality (Zhu et al. 2023). This characteristic not only serves as supportive evidence to the direct application of LLMs in time

EXHIBIT 3

Illustration of Financial Time Series Analysis



series analysis but also paves the way for future multimodal foundation models (Zhang, Gong et al. 2023).

Several notable works have demonstrated the efficacy of LLMs in time series analysis. Pioneering efforts by Zhou et al. (2024) demonstrate the versatility of LLMs across such tasks as forecasting, anomaly detection, classification, and imputation. Using a GPT-2 backbone, they establish the potential for LLMs to process and model time series data effectively. Gruver et al. (2024) further explore the zero-shot capabilities of pretrained LLMs for time series forecasting. Through appropriate tokenization of time series data, they found that LLMs can implicitly understand temporal patterns and generate forecasts without explicit training. Jin, Wang et al. (2023) apply the concept of *reprogramming* to enhance LLM performance in time series analysis. This technique translates time series data into representations more readily understood by LLMs, leading to state-of-the-art forecasting results. Beyond direct LLM applications, researchers are focusing on developing foundation models specifically for time series analysis (Jin, Wen et al. 2023; Liang, Wen et al. 2024). These efforts aim to establish a new paradigm for time series modeling, leveraging the power of techniques in LLMs to capture complex temporal dependencies.

Forecasting

Recent research has explored the utility of LLMs in the domain of financial time series forecasting, demonstrating both the potential and the limitations of these advanced computational tools. This section reviews key studies that have contributed to our understanding of how LLMs can be applied to predict stock market movements and other financial indicators.

LLMs can be directly used for stock forecasting, as shown by Yu et al. (2023). Their research explores NASDAQ-100 stock prediction with LLMs and demonstrates that, by integrating diverse data sources, LLMs not only provide robust predictions but also enhance the explainability. The study emphasizes the importance of instruction-based fine tuning and chain of thought reasoning, which have been shown to significantly improve the performance of LLMs over traditional statistical models in this field. Another way is to integrate LLMs to enhance the other neural networks. Chen, Zheng et al. (2023) introduce a framework that leverages ChatGPT to enhance graph neural networks (GNN) for stock movement prediction. Their approach adeptly extracts evolving network structures from textual data and incorporates these networks into GNNs for predictive tasks. The experimental results indicate that this model consistently outperforms state-of-the-art deep learning-based benchmarks with higher annualized cumulative returns and reduced volatility.

Moreover, LLMs are notable for their capability to be integrated in multimodal data analysis, as discussed in the previous section, which can be crucial when analyzing alternative data. For instance, Wimmer and Rekabsaz (2023) introduce innovative models that leverage both textual and visual data to forecast market movements.

Their research using CLIP-based models shows significant outperformance against established benchmarks in predicting the trends of the German share index. Metrics such as precision, F1 score, balanced accuracy, and others show the effectiveness of these multimodal approaches. Another noteworthy study is the RiskLabs framework, which combines various types of financial data, including textual and vocal information from earnings conference calls, market-related time series data, and contextual news data (Cao et al. 2024). The framework's multistage process starts with extracting and analyzing these data using LLMs, followed by processing time series data to model risk over different timeframes. RiskLabs employs multimodal fusion techniques to combine these varied data features for comprehensive multitask financial risk prediction. The empirical results demonstrate the framework's effectiveness in forecasting both volatility and variance in financial markets, indicating the potential of LLMs in financial risk assessment.

However, the application of LLMs in financial forecasting is not without challenges. Xie, Han, Lai et al. (2023) specifically assess ChatGPT's performance in zero-shot multimodal stock movement prediction tasks and find that it underperforms when compared with both traditional machine learning models and other state-of-the-art techniques. Their findings highlight the necessity for ongoing research to enhance the predictive capabilities of LLMs in complex financial environments. Conversely, Lopez-Lira and Tang (2023) examine how well these models, particularly GPT-4, can predict stock market returns using news headlines as input. Their results indicate that advanced LLMs significantly outperform both traditional models and earlier versions of LLMs. Notably, the models show higher efficacy especially after negative news and for smaller stocks, a phenomenon explained through theories of information diffusion, arbitrage limitations, and investor sophistication. The debate on the effectiveness of LLMs in financial forecasting remains open, with evidence supporting both their limitations and potential.

Though early challenges exist, research reveals considerable promise for LLMs in financial time series forecasting. Explainability, comprehensive understanding of news, and multimodal integration stand out as compelling areas for future investigation and refinement. However, they also mark the challenges and the necessity for further research to fully realize the potential of LLMs in this domain.

Anomaly Detection

Anomaly detection is a fundamental task in various domains, particularly in finance where identifying unusual patterns or outliers is crucial (Chandola, Banerjee, and Kumar 2009). For instance, identifying fraudulent transactions or unusual account activity is a top priority for financial institutions. Anomaly detection algorithms can flag potentially fraudulent behavior, preventing financial losses (Zojaji et al. 2016). Besides, market manipulation schemes, such as pump-and-dump tactics, can be detected through anomaly detection in trading volumes and price patterns (Chen et al. 2019). Anomaly detection is also valuable in risk assessment and mitigation strategies, since anomalies in market trends or macroeconomic indicators can signal potential risks.

Financial time series data, like stock prices, can be highly complex, characterized by volatility, seasonality, and nonlinear relationships. Traditional statistical approaches, though robust, often struggle to encapsulate the full spectrum of these complexities, thereby constraining their anomaly detection capabilities. The development of deep learning has catalyzed a fundamental transformation, offering novel methodologies that hold great promise for this domain (Darban et al. 2022; Crépey et al. 2022). Particularly, LLMs have emerged as a pivotal method, demonstrating remarkable efficacy in anomaly detection across a myriad of tasks, as evidenced by recent scholarly works (Darban et al. 2022; Zhu, Cai et al. 2024). For instance, Park (2024) introduces an LLM-based multi-agent framework that synergizes traditional statistical methods with AI-driven analytics. This innovative fusion is exemplified through an application to the S&P 500 Index, showcasing a marked enhancement in the efficiency, precision, and automation of anomaly detection in financial markets, thereby diminishing the dependency on manual interventions. The integration of LLMs into financial time series anomaly detection will likely become increasingly valuable, which has the potential to not only address the limitations of conventional techniques but also reduce manual processes and enhance algorithmic trading systems that capitalize on market anomalies, paving the way for more sophisticated and automated trading systems.

Other Time Series Tasks

In addition to forecasting and anomaly detection, the capabilities of LLMs offer promising potential within several other domains of financial time series analysis.

Classification. Financial time series can be classified into various categories based on trends, volatility, or other characteristics. LLMs can learn these complex patterns and assign labels accordingly. For instance, they could classify stocks as "growth" or "value," or identify different market regimes—bullish, bearish, and so on (Bosancic, Nie, and Mulvey 2024). LLMs can efficiently classify financial time series data by understanding and predicting patterns that are indicative of specific financial behaviors. This includes the applications of sentiment analysis and anomaly detection that we have already discussed.

Data augmentation. The limited size and variability of financial datasets can sometimes hinder machine learning models. Generative AI offers a path toward data augmentation, which involves generating synthetic data that can be used for training machine learning models, ensuring robustness despite the original limitations of the dataset. A recent paper by Nagy et al. (2023) introduces a generative AI model for end-to-end limit order book modeling, demonstrating the use of a token-level autoregressive generative model to produce realistic order flow in financial markets. This model utilizes structured state-space layers to efficiently handle long sequences of order book states and tokenized messages. The model shows promising performance in approximating data distribution and forecasting mid-price returns, suggesting potential applications in high-frequency financial reinforcement learning. While this work focuses on generative AI rather than directly employing LLMs, its approach and insights are relevant for augmenting financial time series data, highlighting the versatility of generative models in this domain. By simulating various market scenarios, LLMs can help in creating a richer, more diverse dataset that aids in building more accurate predictive models (Ding et al. 2024).

Imputation. Financial time series data often suffers from missing values due to errors or unavailability. Imputation refers to the method of filling in missing or incomplete data points in financial time series. LLMs have a good potential to fill these missing values based on their superior generative capability (Zhao, Zhou et al. 2023). This is particularly useful in maintaining the quality and continuity of financial data analysis. Accurate imputation helps in avoiding biases or inaccuracies that might occur due to gaps in the data, thus ensuring more reliable financial assessments and forecasts.

In summary, LLMs demonstrate significant potential in financial time series analysis, offering capabilities in forecasting, anomaly detection, pattern classification, data augmentation, imputation, and more. Their ability to process and understand complex financial data opens avenues for novel approaches to market analysis. As LLM research progresses, we can anticipate continued advancements in the application of these models within the financial time series domain.

FINANCIAL REASONING

Another key application of LLMs in finance is to support financial reasoning. As previously discussed, LLMs are capable of processing and synthesizing vast amounts of financial data from various sources, including market reports, financial news, and historical pricing data. This comprehensive understanding of the financial landscape and market dynamics may enable LLMs to support strategic financial planning, generate investment recommendations, provide advisory services, and assist in financial decision making (see Exhibit 4).

The use of LLMs in financial reasoning offers several key advantages. First, they can *enhance data analysis* by processing vast amounts of financial information, identifying patterns and trends that help inform better decision-making. Second, LLMs can be used for *predictive modeling*, allowing them to forecast market conditions and asset performance, which may lead to robust investment recommendations. Additionally, LLMs could offer *personalized advisory services*. They can analyze a person's or organization's financial situation, goals, and risk tolerance to provide customized advice. Another benefit could be *real-time monitoring and alerts*, where LLMs can monitor financial market trends and news, providing timely updates and alerts to help users adjust their strategies as needed. Moreover, LLMs may *improve accessibility and engagement*. By integrating these models into user-friendly interfaces, such as chatbots, financial planning and advisory tasks become more accessible and engaging, where individuals can take control of their own financial well-being.

In this section, we will explore these applications through the literature, potentially inspiring further innovations.

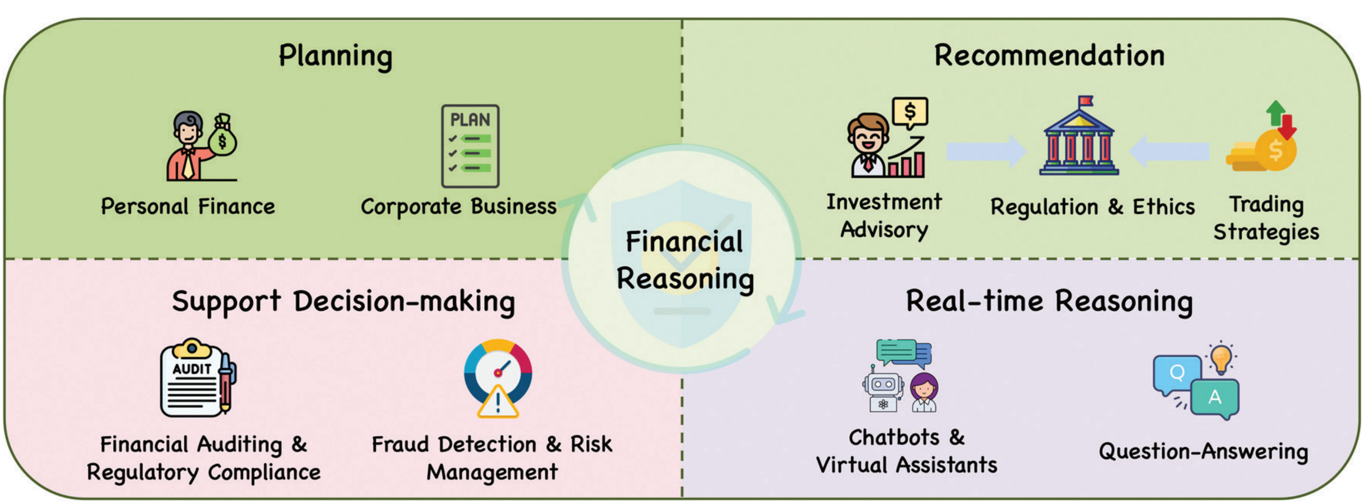
Planning

Financial planning involves setting financial goals, assessing current financial situations, and devising strategies to achieve those goals. This process includes analyzing income, expenses, investments, and risk management to create a comprehensive plan for long-term financial stability and growth.

In a corporate context, LLMs can be utilized to support various aspects of financial planning. For instance, LLMs can analyze market trends and competitor data to

EXHIBIT 4

Illustration of Various Financial Reasoning Tasks



help organizations develop business strategies. Nguyen and Tulabandhula (2023) examine the use of generative AI models, such as GPT-4 and other transformer-based models, for business strategy development. By employing named entity recognition (NER) and zero-shot classifiers (ZSC) to automatically extract and classify relationships among companies, they created dynamic signed business networks that reflect the competitive and collaborative market landscape. This method provides business stakeholders with insights into market conditions and supports strategic decision-making.

Moreover, LLMs can streamline financial planning processes, as demonstrated by Ludwig and Bennetts (2023). By integrating ChatGPT into financial planning practices, they illustrate how financial planners could leverage this AI model to enhance client communication and provide immediate, semi-personalized responses to common financial concerns, such as preparing for economic recessions. They also highlight ChatGPT's role in client education and its ability to simplify complex financial concepts for better understanding. Despite these benefits, the authors emphasize the need for human oversight to ensure the accuracy and quality of the advice provided, addressing potential limitations of the models.

In personal financial planning, LLMs can help individuals create customized strategies for long-term financial well-being. A recent study by Lakkaraju, Vuruma et al. (2023) evaluates the performance of LLM-based chatbots, ChatGPT and Bard, in providing personal financial advice. The study covers various aspects of personal finance, including decisions related to bank accounts, credit cards, and certificates of deposits (CDs). It assesses how these models handle complex financial interactions and make recommendations across different languages and dialects, such as English, African American Vernacular English, and Telugu. Their findings reveal that while ChatGPT often provides more personalized and accurate responses, both models face challenges, including mathematical errors, lack of visual aids to support explanations, and difficulties in processing non-English queries effectively. The paper emphasizes the need for improvements in these LLMs to enhance their reliability and inclusivity when applied to financial planning.

Additionally, LLMs can optimize budgeting strategies by incorporating AI-driven recommendations into individual and household financial models. In de Zarzà et al. (2023), the authors present an optimization framework for individual budget allocation to maximize savings and extend this approach to household finances, addressing the complexities of multiple incomes and shared expenses. In high-net-worth contexts, LLMs can also be used to simulate various tax scenarios, identify optimal tax strategies, and provide proactive advice based on changing tax law to minimize tax liabilities and maximize financial growth (Fava 2023).

The integration of LLMs in financial planning has the potential to transform how individuals and businesses approach their financial objectives. By leveraging the data processing and analysis capabilities of LLMs, financial planning can become more efficient, accurate, and personalized. As research and development in this field continue to progress, LLMs are poised to become vital tools in the financial planning environment, allowing users to make educated and strategic decisions. The examples discussed in this section highlight the diverse range of applications and the potential for LLMs to revolutionize financial planning practices for both corporate and personal contexts.

Recommendation

LLMs are revolutionizing investment recommendations and wealth management by analyzing financial data, forecasting market trends, and optimizing portfolios. They provide personalized advice based on individual risk profiles and preferences, which improves robo-advisors and investment strategies. However, the integration of LLMs in wealth management needs regulatory frameworks to assure fairness, effectiveness, and informed decision-making in conjunction with human expertise.

LLM in investment advisory. LLMs play a crucial role in enhancing the capabilities of robo-advisors by providing personalized and automated investment recommendations. For example, Huang et al. (2024) highlight the effectiveness of platforms, such as Wealthfront and Betterment, that employ AI algorithms to deliver customized asset management plans aimed at optimizing investment returns based on individual user profiles. The study emphasizes the importance of consistent use, transparency, and user-centric design in maximizing the benefits of intelligent advisors. To build user trust and improve the overall effectiveness of robo-advisors, the authors recommend focusing on key areas such as enhancing transparency, designing intuitive user interfaces, and offering tailored financial services for individual needs.

Similarly, Lu, Huang, and Li (2023) explore the potential of ChatGPT in generating investment portfolio recommendations. Using textual data from *The Wall Street Journal* and Chinese policy announcements, the researchers evaluate ChatGPT's ability to generate portfolios that outperform the market. Through fine-tuning and performance measurements, the study demonstrates that ChatGPT can achieve a monthly three-factor alpha of up to 3%, particularly when analyzing policy-related news. They highlight the importance of model parameters, such as the "temperature" setting, in influencing the recommendations' creativity and accuracy, indicating that generative AI, with appropriate tuning, can be a valuable tool for financial advisors.

Another development in the field is the Cogniwealth system, introduced by Ramyadevi and Sasidharan (2024). This platform utilizes the Llama 2 model as a financial advisor. The system leverages NLP and machine learning techniques to assist both professional fund researchers and laymen investors by providing personalized investment recommendations and financial insights. Cogniwealth's ability to handle user-provided data and deliver human-like responses through an intuitive interface ensures high levels of adaptability, user-friendliness, and engagement.

Impact on investment strategies. LLMs are transforming the landscape of investment strategies, offering the potential to deliver more accurate, diverse, and accessible investment advice. A prime example of this is demonstrated in Ko and Lee's (2024) study, where they showcase ChatGPT's ability to construct portfolios with superior diversity and performance compared with randomly selected ones. This finding highlights the potential of LLMs to serve as valuable advisory tools for both professional portfolio managers and individual investors, democratizing access to advanced investment strategies.

LLMs can also impact the development of algorithmic trading strategies by automating the creation of accurate and executable code for technical indicators. The study conducted by Noguer i Alonso and Dupouy (2024) compares the capabilities of various LLMs, such as GPT-4-Turbo, Gemini-Pro, Mistral, Llama 2, and Codellama, in generating code that runs correctly and matches baseline implementations. The study emphasizes the importance of well-designed prompts and the models' ability to handle complex financial calculations for successful code generation.

Recently, Kim, Muhn, and Nikolaev (2024) investigate the capability of an LLM, specifically GPT-4 Turbo, to perform financial statement analysis comparable to that of professional human analysts. By providing standardized and anonymous financial statements, the study examines the model's ability to predict future earnings without any narrative or industry-specific context. The findings reveal that the LLM not only outperforms human analysts in predicting earnings changes, particularly in challenging scenarios, but also matches the performance of specialized state-of-the-art machine learning models. The authors claim that the model's predictions derive not from its training memory but from generating useful narrative insights about a company's future performance, thus eliminating look-ahead bias. To address this bias, the research design uses a consistent anonymized format for financial statements across firms, making it virtually impossible for the model to infer a firm's identity. Additionally, the statements do not contain any dates and use relative years, mitigating concerns about the model leveraging macroeconomic trends from specific years. Furthermore, trading strategies based on the LLM's predictions demonstrate higher Sharpe ratios and alphas compared to those based on other models.

Another promising application of LLMs in investment strategies is the analysis of annual reports to extract valuable insights, thereby enhancing stock investment strategies. Gupta (2023) introduces a framework that utilizes GPT-3.5 to streamline the process of analyzing comprehensive 10-K filings of a company. By combining the generated insights with historical stock data, the study demonstrates that machine learning models trained on these LLM-generated features can outperform traditional market benchmarks, such as the S&P 500. This approach highlights the potential of integrating LLMs with historical data to improve the accuracy of stock predictions and enhance investment strategies.

Moreover, Zhang, Yoshie, and Huang (2024) introduce BreakGPT for detecting financial breakouts. BreakGPT's multistage structure improves the accuracy and stability of detecting true and false breakouts in financial markets by systematically analyzing price movements and order flows. The model's superior performance compared with ChatGPT-3.5 and ChatGPT-4 makes it a valuable tool for traders and investors in detecting financial breakouts.

However, despite these promising developments, Chuang and Yang (2022) raise an important concern regarding the implicit biases present in pretrained language models, such as BERT and FinBERT. The study reveals that these models exhibit significant biases toward certain stocks and industry sectors, which can influence the quality and fairness of investment recommendations. They emphasize the need for awareness and mitigation of such biases in financial decision-making systems to ensure more reliable and fair investment advice. This research highlights the importance of careful model training and evaluation in financial contexts to develop robust and accountable financial advisory systems.

Regulatory and ethical considerations. The application of LLMs in financial advisory services has raised significant regulatory and ethical concerns. Caspi, Felber, and Gillis (2023) examine the regulatory landscape, highlighting key concerns such as maintaining fiduciary duties, ensuring transparency, and preventing conflicts of interest. They discuss potential regulatory strategies to address the challenges posed by generative AI, emphasizing the need for effective regulation that balances innovation with consumer protection. Moreover, Niszczota and Abbas (2023) investigate the financial literacy of GPT models, revealing GPT-4's near-perfect score on financial literacy tests. However, they also find that individuals with lower financial knowledge tend to rely more heavily on GPT's advice.

Lakkaraju, Jones et al. (2023) also compare the effectiveness and fairness of LLM-based chatbots (ChatGPT and Bard) with a rule-based chatbot (SafeFinance) in providing personal financial advice. They find that while ChatGPT and Bard generate fluent responses, they exhibit inconsistencies and biases across different user groups and languages. In contrast, SafeFinance provides reliable answers, albeit with limited generalization. The study demonstrates the need for improvements in LLM-based systems to ensure fairness and accuracy in financial advisement.

While LLMs have demonstrated potential in transforming financial advisory services, their application raises important regulatory and ethical considerations. Effective regulation should balance innovation with consumer protection, while educating users about the limitations and potential biases of AI-driven financial recommendations is essential to promote informed decision-making.

Support Decision-Making

Operational risk management and compliance are critical components in the financial sector, as they help safeguard the integrity of financial institutions, protect consumers, and maintain the stability of the entire financial system. However, the increasing complexity of financial products, ever-changing regulations, and the constant threat of fraudulent activities pose significant challenges for financial institutions. LLMs are emerging as powerful tools that enhance these processes by providing sophisticated analytical capabilities. By leveraging LLMs, financial institutions can improve the accuracy of audits, streamline compliance verification, and detect inconsistencies more efficiently. This enables financial institutions to make informed decisions in such critical areas as financial auditing and regulatory compliance, and fraud detection and risk management, ultimately enhancing their operational resilience and ensuring compliance with regulatory requirements.

Financial auditing and regulatory compliance. Financial auditing involves the systematic examination of financial records and statements to ensure accuracy and compliance with regulations. LLMs are increasingly being used to enhance these processes by improving the accuracy and efficiency of text matching and regulatory interpretation (Berger et al. 2023). A study conducted by Hillebrand et al. (2023) introduces ZeroShotALI, which stands for Zero-Shot Automated List Inspection. It combines GPT-4 and a domain-specific SentenceBERT model to enhance the matching of text segments from financial reports with specific legal requirements. This system significantly improves the efficiency and accuracy of financial audits compared to traditional methods.

Moreover, another study conducted by Cao and Feinstein (2024) examines the use of LLMs (such as GPT-4, GPT-3.5, Claude-3-Opus, Gemini-1.5-Pro) for interpreting complex financial regulations, specifically focusing on Basel III capital requirements. Effective prompt design and document loading methods guide LLMs in translating regulatory texts into a concise mathematical framework, aiming to significantly enhance regulatory interpretation accuracy.

In addition, by analyzing firms' public narrative disclosures with GPT-4, Choi and Kim (2024) develop a novel measure of tax audit periods at the firm level. Their measure shows high conformity with data released by the US Internal Revenue Service and reveals that tax audits lead to reduced tax avoidance, decreased capital investments, and increased stock volatility.

LLMs have shown potential in uncovering inconsistencies and contradictions in financial reports. A study conducted by Deußer et al. (2023) develops an innovative approach to identifying discrepancies in financial reports by leveraging the power of LLMs such as GPT-4 and Llama. The study employs embedding-based paragraph clustering to efficiently detect contradictions across various datasets, including both annotated and unannotated financial reports. By utilizing sentence-pair data, document-level data, and intelligent bucketing systems, the researchers optimize the query process for the LLMs, enabling them to effectively pinpoint inconsistencies and contradictions. The results of this study demonstrate significant enhancements in the accuracy and efficiency of financial audits, ultimately reducing the time and effort required to conduct thorough and reliable financial report audits.

Fraud detection and risk management. Fraud detection and risk management are critical components of maintaining financial integrity and stability. LLMs offer advanced capabilities to detect fraudulent activities and manage risks through sophisticated data analysis and pattern recognition. A study conducted by Feng et al. (2023) highlights the potential of LLMs to revolutionize credit scoring and risk assessment. By instruction tuning, LLMs can match or surpass traditional credit scoring models, leading to more inclusive and comprehensive evaluations. However, the study also emphasizes the need to address biases within LLMs to ensure fair financial decision making.

Furthermore, Cao et al. (2024) present a novel framework named RiskLabs that leverages LLMs to predict financial risk by integrating data from various sources. By processing and fusing features from diverse data types, including textual and vocal information from earnings conference calls (ECCs), market-related time series data, and contextual news data surrounding ECC release dates, RiskLabs outperforms traditional methods and existing models in forecasting financial risks and provides a more comprehensive understanding of market dynamics.

Several papers explore the application of LLMs in fraud detection. Zhao, Zhu et al. (2023) introduce an innovative GPT-based model for identifying fraudulent activities in payment systems, which excels in capturing detailed behavioral sequences through temporal and contextual analysis. Yang et al. (2023) introduce the FinChain-BERT model, which enhances fraud detection accuracy by focusing on key financial terms and optimizing model performance. Similarly, Bhattacharya and Mickovic (2024) demonstrate the effectiveness of the BERT model in detecting accounting fraud in financial reports by fine-tuning the model on management discussion and analysis sections of annual 10-K reports from the US Securities and Exchange Commission (SEC) database, outperforming existing benchmark models.

While LLMs have shown great potential in fraud detection and risk management, it is crucial to acknowledge and address the inherent biases that may exist within these models. Biases in LLMs can lead to unfair and discriminatory practices in financial decision-making. Ongoing research and development efforts are necessary to mitigate these biases and ensure the responsible and ethical deployment of LLMs in the financial sector.

Real-Time Reasoning

Real-time reasoning enables instant and dynamic interactions between users and AI-powered systems. By leveraging the vast knowledge and understanding of LLMs, financial institutions can deploy chatbots, virtual assistants, and question-answering systems that provide accurate, relevant, and timely information to customers and stakeholders. These real-time applications streamline customer support, simplify complex financial transactions, and offer immediate access to financial insights and advice.

Chatbots and virtual assistants. Chatbots and virtual assistants are changing the way financial institutions interact with customers and streamline internal processes. By leveraging the capabilities of LLMs, these AI-driven tools can further provide more personalized and effective assistance, thereby enhancing customer satisfaction and boosting organizational efficiency. For instance, Aggarwal, Mehra, and Mitra (2023) present a multipurpose NLP chatbot that incorporates LLM models, including ChatGPT, BERT, and DistilBERT. The proposed system incorporates emotion recognition, multilingual support, and voice conversion. The chatbot demonstrates exceptional performance in providing personalized financial advice, understanding and responding to human emotions, and maintaining functionality in offline modes.

In another study, Yue and Au (2023) introduce GPTQuant, a conversational AI chatbot designed to facilitate investment research. GPTQuant leverages few-shot learning and LangChain's integration to generate Python code for backtesting and strategy analysis. The chatbot uses prompt templates to activate GPT-3's capabilities, demonstrating efficacy in portfolio construction, rebalancing, and factor score querying.

Lastly, Yadav et al. (2024) introduce a virtual assistant that uses LLMs to enhance the financial reconciliation process. The assistant automates the generation of SQL queries from natural language inputs, streamlining and expediting reconciliation, research, and validation processes for accountants. Utilizing a retrieve-and-refine strategy with retrieval-augmented generation (RAG) and few-shot prompting, the virtual assistant achieved 95% accuracy in generating correct SQL queries for real-world questions related to account reconciliation. This integration of LLMs significantly improves the accuracy and efficiency of generating SQL queries, demonstrating the potential of LLMs to automate repetitive and time-consuming tasks in financial reconciliation.

Question answering. Question-answering systems powered by LLMs have shown remarkable progress in understanding and responding to complex queries related to financial documents. Recent studies have focused on enhancing the numerical reasoning capabilities of these systems, enabling them to handle multistep calculations and extract relevant information from various data sources. For example, Arun et al. (2023) develop a pipeline utilizing fine-tuned LLMs, such as Llama-2-7B and T5, to analyze financial reports and answer numerical reasoning questions. By extracting and serializing tables from PDFs, generating embeddings, and training on the FinQA dataset, the authors demonstrated the potential for real-time analysis of financial reports. The study concludes that with appropriate fine-tuning and methodologies, LLMs could significantly enhance the efficiency and accuracy of financial data analysis, enabling swift and informed decision-making in dynamic market environments through rapid extraction and interpretation of crucial data points.

Furthermore, Phogat et al. (2023) introduce zero-shot prompting techniques (ZS-FinPYT and ZS-FinDSL) for LLMs including GPT-3, GPT-3.5-turbo, and GPT-4 to perform complex numerical reasoning over financial documents. By encoding reasoning into Python/DSL(domain-specific languages) programs, these techniques mitigate arithmetic limitations. Evaluations on datasets such as FinQA, ConvFinQA, and TATQA demonstrate superior performance compared with baselines, particularly in table/text data, multistep reasoning, and numerical questions.

In a related study, Srivastava, Malik, and Ganu (2024) investigate the mathematical reasoning capabilities of LLMs on financial documents. They introduce a novel prompting strategy, EEDP (elicit-extract-decompose-predict), designed to enhance LLM performance in scenarios requiring multistep numerical reasoning. Extensive experimentation with multiple LLMs across financial datasets reveals that EEDP outperforms baseline strategies like direct prompting, chain of thought (CoT), and program of thoughts (PoT). The study highlights the potential of structured prompting strategies in improving LLM performance for complex reasoning tasks and identified common error types, emphasizing the need for precise information extraction.

Xue et al. (2023) propose a cutting-edge dialogue system designed specifically for the finance sector, named WeaverBird. It leverages a LLM with GPT architecture fine-tuned on extensive financial corpora. This enables WeaverBird to understand and provide informed responses to complex financial queries, such as investment strategies during inflation. The system's performance is further enhanced by integrating a local knowledge base and search engine, allowing it to retrieve relevant information and generate responses conditioned on Web search results, complete with proper source references for enhanced credibility. Comparative evaluations across a broad spectrum of financial question-answering tasks demonstrate WeaverBird's superior performance compared with other models, positioning it as a powerful tool for financial dialogue and decision support.

AGENT-BASED MODELING

Agent-based modeling (ABM) represents a significant advancement in simulating complex systems, particularly in finance. The core principle of ABM involves creating autonomous agents that interact within a defined environment, allowing the emergence of complex phenomena from the bottom up. Unlike traditional models that assume uniform behavior among agents and equilibrium states, ABM captures the diversity of behaviors and adaptive strategies that characterize real-world financial markets. This flexibility makes ABM a powerful tool for understanding market dynamics, investor behavior, and the impact of various external factors on financial systems (as illustrated in Exhibit 5).

In recent years, the integration of LLMs with agent-based modeling has opened new avenues for research and application (Guo et al. 2024; Xi et al. 2023; Ma et al. 2024). With their advanced NLP capabilities, LLMs enhance the cognitive functions of agents, allowing them to interpret and react to vast amounts of unstructured data such as financial news, reports, and social media posts. This synergy between LLMs and ABM leads to more realistic and adaptive simulations, which are crucial for developing robust trading and investment strategies (Zhang, Mao et al. 2024).

Traditional applications of ABM in finance have focused on modeling the interactions among different types of market participants, such as institutional investors, individual traders, and regulatory bodies (Epstein 1999). These models have been used to study the impact of regulatory changes, market shocks, and behavioral biases on market dynamics. For instance, agent-based models have been employed to simulate the effects of high-frequency trading, the propagation of financial crises, and the formation of asset bubbles. The addition of LLMs to these models further enhances their predictive power and accuracy by enabling agents to process and respond to real-time information in a manner similar to human analysts.

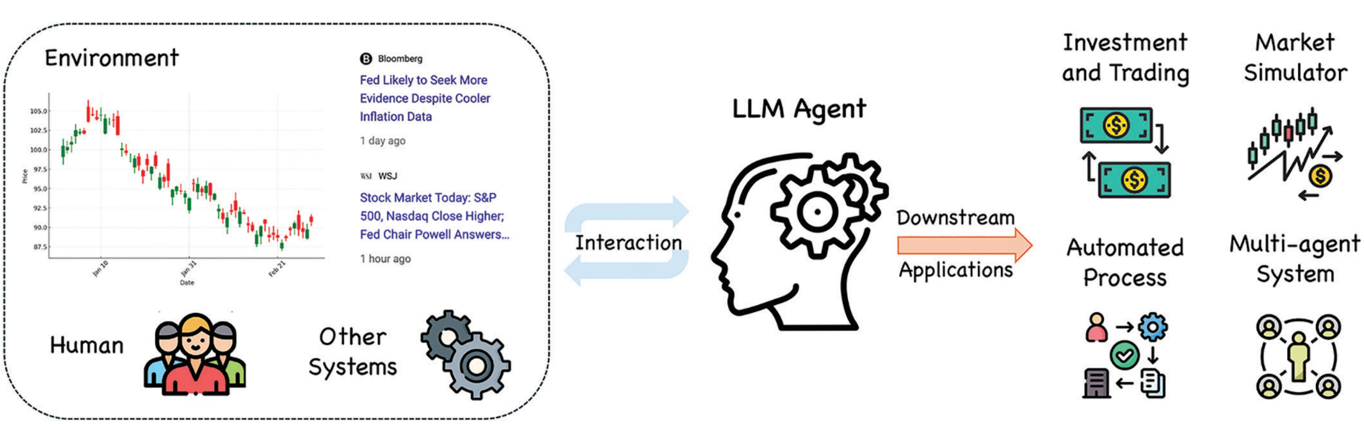
In this section, we explore the integration of LLMs with agent-based modeling in various contexts. We discuss how LLM-based trading and investment agents enhance decision-making and strategy formulation. We also examine the use of LLMs in simulating markets and economic activities, highlighting their impact on policy analysis and market predictions. Additionally, we review the role of multi-agent systems in improving financial process automation and monitoring, emphasizing the potential of these advanced models in revolutionizing financial analysis and strategy development.

Trading and Investments

The financial markets are dynamic and complex, requiring advanced tools to navigate effectively. LLMs have proven to be powerful allies in this domain by enabling the creation of intelligent trading agents that can process vast amounts of data and execute

EXHIBIT 5

Illustration of Financial Tasks Related to Agent-Based Modeling



trades with high precision. These agents leverage LLMs' NLP capabilities to interpret and synthesize financial news, market reports, and historical data, significantly improving market predictions and trading strategies. **StockAgent** (Zhang, Liu et al. 2024), for instance, explores the potential of AI-driven trading systems to simulate and analyze stock market behaviors under various external influences. It is a multi-agent system powered by LLMs designed to mimic real investor behaviors and assess the impact of external factors such as macroeconomic events, policy changes, and financial reports on trading activities. The study finds that different LLMs, such as GPT-3.5 Turbo and Gemini, exhibit distinct trading behaviors and preferences, with GPT agents showing more diverse and independent trading styles compared with the more homogeneous and trend-following behavior of Gemini agents. This variation suggests that LLM-based systems can offer personalized investment strategies and insights. The research also highlights that removing financial information or communication channels like BBS (Bulletin Board System) can significantly alter trading behaviors and market dynamics, underscoring the complexity and interdependence of factors influencing stock trading.

A notable advancement in LLM applications is integrating multimodal data—textual, numerical, and visual—into trading agents. **FinAgent** (Zhang, Zhao et al. 2024) exemplifies this by combining these data types to support quantitative and high-frequency trading, including stocks and cryptocurrencies. Its diversified memory retrieval system and tool augmentation features enable FinAgent to interact with various data sources and tools, enhancing adaptability and performance in dynamic trading environments.

LLM-based trading agents excel in continuous learning and adaptation as well. **FINMEM** (Yu et al. 2024) introduces a layered memory and character design, enhancing the agent's ability to process hierarchical financial data and convert insights into trading decisions. The memory module of FINMEM, inspired by human cognitive processes, includes working memory and layered long-term memory components. This design allows FINMEM to categorize and prioritize information based on its relevance and timeliness, retaining critical insights longer and enabling agile responses to new investment cues. Through real-world testing and continuous learning, FINMEM evolves its trading strategies, demonstrating improved decision making and adaptability in volatile financial environments. Similarly, **QuantAgent** (Wang, Yuan et al. 2024) focuses on self-improvement through a two-layer loop system. The inner loop refines responses using a knowledge base, while the outer loop involves real-world testing and knowledge enhancement. This iterative approach enables QuantAgent to autonomously extract financial signals and uncover viable trading opportunities, showcasing LLMs' dynamic potential.

Integrating human expertise with AI capabilities is another significant advancement. The **Alpha-GPT** series, including Alpha-GPT (Wang, Yuan et al. 2023) and Alpha-GPT 2.0 (Yuan, Wang, and Guo 2024), emphasizes human-AI interaction in the alpha mining process. Alpha-GPT 2.0 further introduces a human-in-the-loop framework for iterative refinement of investment strategies. These agents interpret trading ideas and translate them into effective strategies, providing insightful and actionable alphas. By leveraging both human expertise and AI capabilities, this approach enhances the efficiency and creativity of the alpha mining process, leading to more effective investment decisions.

Simulating Markets and Economic Activities

Simulating markets and economic activities has long been a critical aspect of financial research and policy analysis. Traditional simulators, typically grounded in econometric models and system dynamics, have been the cornerstone of this effort. These simulators rely on historical data and established economic theories to predict future market behaviors. For instance, models like the vector autoregression (VAR) model and the dynamic stochastic general equilibrium (DSGE) model are widely used for economic forecasting and policy analysis (Sims 1980; Smets and Wouters 2003). While they offer a structured and mathematically rigorous approach, traditional simulators often struggle with the complexity and dynamism inherent in real-world economic systems. They are generally static, assuming rational behavior and equilibrium, which can limit their accuracy and adaptability to unforeseen economic shocks or behavioral intricacies.

In contrast, agent-based simulators represent a significant advancement in the simulation of economic activities. These models consist of autonomous agents, each with distinct behaviors and decision-making processes. These agents interact within a defined environment, allowing for the emergence of complex macroeconomic phenomena from the bottom up. The primary advantage of ABM lies in their flexibility and ability to model heterogeneous agents with varying strategies and interactions. This approach can capture the nonlinear dynamics of markets, such as feedback loops, market sentiments, and adaptive behaviors (Tesfatsion 2006).

However, agent-based simulators are not without their challenges. One significant drawback is the computational complexity, as simulating numerous agents with intricate interactions demands substantial processing power. Additionally, the development of realistic agent behaviors and interaction rules requires deep domain expertise and can be time-consuming. Moreover, while agent-based simulators can model emergent phenomena, the validation of these models against real-world data remains a challenging task, often requiring extensive calibration and sensitivity analysis (Farmer and Foley 2009).

The integration of LLMs with agent-based simulators represents a cutting-edge development in economic simulations. With their advanced NLP capabilities, LLMs can enhance the perception, reflection, and decision-making processes of agents within simulators. This hybrid approach leverages the strengths of both technologies: the detailed and adaptive behaviors modeled by agent-based simulators and the comprehensive data processing and learning capabilities of LLMs.

Research by Li, Gao, Li, and Liao (2023) exemplifies the potential of this integration by demonstrating the ability to simulate complex macroeconomic activities. Their study, **EconAgent**, shows how LLM-empowered agents can realistically model economic activities by processing economic data through advanced mechanisms. These agents can simulate human-like decision-making processes, providing a comprehensive understanding of how different economic factors interact. This enables more accurate predictions of economic trends and the effects of policy changes. Equipped with layered memory systems, these agents can adapt their strategies based on real-time data inputs and historical analysis, making them highly effective for forecasting and policy simulation.

Similarly, Horton (2023) explores the use of LLMs as computational models for economic simulations. By endowing LLMs with preferences and decision-making frameworks, their approach allows the simulation of human-like economic behavior. These simulations are particularly valuable for social science experiments and exploring economic scenarios, providing insights that can inform policy and strategy. The study introduces "Homo Silicus" agents, designed to emulate human economic agents by incorporating principles of behavioral economics. This enables the agents to make decisions based on a mix of rational analysis and emotional factors, providing a more realistic simulation of economic activities and market behaviors.

Furthermore, Zhao, Wang et al. (2023) investigate the competitive behaviors of LLM-based agents in a simulated environment, demonstrating how competition among agents can lead to the emergence of innovative strategies and enhanced performance. They propose **CompeteAI**, a framework that simulates a virtual town where restaurant agents compete for customers, revealing how competition drives agents to continually adapt and improve their strategies, aligning with established sociological and economic theories.

The evolution from traditional simulators to agent-based models and now to LLM-empowered agents marks a significant stride in the field of economic simulation. The integration of LLMs with ABM offers a promising avenue for more realistic and adaptive modeling of economic activities, capturing the complex interplay of factors that drive markets and economies. This hybrid approach not only enhances our understanding of economic dynamics but also provides a powerful tool for forecasting and policy analysis.

Automated Financial Processes

The integration of LLMs into financial processes has reformed the way financial tasks are automated, offering enhanced capabilities for workflow generation and strategic planning. These applications streamline operations and provide robust solutions for complex financial tasks.

One notable application is **FlowMind** (Zeng et al. 2023), which presents an innovative approach to automating financial workflows using LLMs. FlowMind leverages the capabilities of models like GPT to generate workflows on the fly, addressing the limitations of traditional robotic process automation that relies on predefined tasks. The system uses a structured lecture recipe to ground LLM reasoning with reliable APIs, mitigating issues such as hallucinations and ensuring data privacy by avoiding direct interaction with proprietary code. FlowMind includes a feedback loop that allows users to inspect high-level descriptions of the generated workflows and provide adjustments, enhancing the system's adaptability. This approach is demonstrated using the NCEN-QA dataset, a benchmark for evaluating workflow generation in financial question-answering tasks, where FlowMind significantly outperforms traditional methods. This framework showcases the potential of LLMs to automate complex, spontaneous tasks in financial services while maintaining data integrity and security.

Another application is **AUCARENA** (Chen, Yuan et al. 2023), which evaluates strategic planning and execution in auction environments to assess the strategic reasoning capabilities of LLM agents. In the ascending-price auctions, LLM agents like GPT-4 compete, managing budgets and adapting strategies in real time. Utilizing a belief-desire-intention model, agents update beliefs, adjust desires, and replan based on auction developments. This setup allows for a detailed analysis of how LLM agents manage resources, adhere to goals, and adapt to new information in competitive contexts. The study shows that LLM agents, especially GPT-4, are effective in strategic planning and resource management, though sometimes outperformed by simpler methods, highlighting areas for further improvement in LLM design. AUCARENA demonstrates the potential of LLMs to enhance decision-making processes in complex, competitive scenarios.

Multi-Agent Systems

The use of multi-agent systems in financial analysis leverages the strengths of LLMs to enhance the robustness and accuracy of financial strategies. Multi-agent systems improve trading performance by simulating various agent interactions and providing a more comprehensive analysis of the tasks. **TradingGPT** (Li, Yu, Li, Chen, and Khashanah 2023) exemplifies this approach with its innovative multi-agent framework designed for financial trading. It organizes memory into three distinct layers: short term, medium term, and long term, each governed by a custom decay mechanism that matches human cognitive processes. In TradingGPT, agents can engage in inter-agent communication and debate, enhancing their decision-making capabilities. Each agent is equipped with individualized trading characters, such as risk-seeking, risk-neutral, and risk-averse profiles, which enrich the diversity of perspectives and improve decision-making robustness. By leveraging layered memory processing and consistent information exchange, this framework demonstrates augmented adaptability to historical trades and real-time market cues, significantly enhancing automated trading outcomes.

Aside from trading tasks, **SocraPlan** (Tsao and Chang 2023) leverages multi-agent reasoning with LLMs for effective corporate planning. This framework conducts comprehensive market research, customer profiling, product usage analysis, and sales strategy formulation. By combining human insights with AI capabilities, SocraPlan enhances corporate planning, enabling businesses to devise strategies that are both innovative and grounded in detailed market analysis. SocraPlan employs a multi-agent architecture where each agent specializes in a different aspect of corporate planning, such as competitive analysis, customer segmentation, or trend forecasting. These specialized agents collaborate to provide a holistic view of the market, which helps businesses make informed strategic decisions.

Multi-agent systems also benefit in analyzing financial sentiments or textual information, which is a critical component of market analysis and strategy formulation as we have discussed previously. An example is **HAD** (Xing 2024) or heterogeneous agent discussion, employing specialized agents focused on different types of errors common in FSA. This framework ensures that each of the agents focuses on particular errors, such as sarcasm, aspect mismatches, and temporal expressions, making the system robust against common pitfalls in sentiment analysis. The HAD framework has shown significant improvements in accuracy and F-1 scores across multiple datasets, proving its efficacy in refining sentiment analysis for financial texts. Another example is by Wan et al. (2024), who introduce a multi-agent framework that automates the verification of information between loan applications and bank statements, powered by both open-sourced models such as Llama 3 and close-sourced models such as GPT-4. Despite higher operational costs, this approach is more economical and faster than manual reviews, offering a reliable solution for structured finance auditing and compliance.

Moreover, multi-agent systems can be used for monitoring and anomaly detection in financial markets. Park (2024) introduces a sophisticated multi-agent framework designed to improve the validation and interpretation of financial data anomalies. The framework employs a network of specialized LLM agents, each focusing on distinct tasks such as data conversion, web-based expert analysis, utilization of institutional knowledge, cross checking, and report consolidation. This collaborative approach enhances the efficiency and accuracy of anomaly detection, reducing the need for manual verification. By applying this framework to the S&P 500, the study demonstrates significant improvements in anomaly detection, showing that LLM-based agents can autonomously and accurately identify and interpret anomalies in financial market data, thereby supporting more effective financial market monitoring and decision-making.

Besides the multi-agent systems, an agent can interact with itself in an autonomous way as well (Su et al. 2022). The self-reflective LLM framework or **SEP** (Koa et al. 2024), named for “summarize-explain-predict,” addresses this need by enabling the generation of explainable stock predictions. SEP combines verbal self-reflective agents with proximal policy optimization (PPO) to provide autonomous and explainable predictions. This framework allows agents to self-reflect on their decision-making processes, ensuring that the predictions are not only accurate but also interpretable. By enhancing the explainability of stock predictions, SEP improves accuracy, transparency, and trustworthiness among investors and analysts.

In summary, the integration of LLMs into agent-based modeling in finance offers significant advancements in trading, investment, financial analysis, and economic simulation. These applications demonstrate the versatility and effectiveness of LLMs in enhancing decision-making, strategy formulation, and market analysis. Future research in this area promises to further refine these systems, improving their accuracy, efficiency, trustworthiness, and adaptability in the ever-evolving financial landscape (Sharma et al. 2024; Zhang, Bo et al. 2024; Hua et al. 2024).

DATASETS

The datasets used in this article cover a wide range of financial domains and tasks. These datasets are crucial for training and evaluating models on specific financial tasks such as sentiment analysis, question answering, relation extraction, and numerical reasoning. Several widely used datasets are as follows:

- *Financial PhraseBank (**FPB**; Malo et al. 2014).* This is a dataset consisting of financial phrases annotated with sentiment labels. It is widely used for sentiment analysis in financial contexts due to its detailed and domain-specific annotations.
- *Financial Question Answering and Opinion Mining (**FiQA**; Maia et al. 2018).* This dataset focuses on aspect-based sentiment analysis and opinion-based question answering. It includes financial news headlines and microblogs, annotated for sentiment and aspect categories. The dataset is designed to challenge models with tasks that require fine-grained sentiment and opinion extraction from financial texts.
- ***FinQA** (Chen et al. 2021).* A dataset designed for numerical reasoning over financial data. FinQA includes questions that require understanding and manipulating numerical information from financial reports. It emphasizes the need for models to perform complex reasoning tasks involving financial metrics and calculations.

Other datasets, such as **ECTSum** (Mukherjee et al. 2022), **FiNER** (Shah, Paturi, and Chava 2023), **FinRED** (Sharma et al. 2022), **REFinD** (Kaur et al. 2023), **FinSBD** (Au, Ait-Azzi, and Kang 2021), and **CFLUE** (Zhu, Li et al. 2024), contribute to various specific financial NLP tasks. These include earnings call summarization, named entity recognition, relation extraction, and financial language understanding evaluations. Collectively, these datasets provide a robust foundation for developing and benchmarking LLMs in financial applications.

BENCHMARKS AND CODE

We outline the comprehensive benchmark used to assess LLM performance in the financial domain as shown in Exhibit 6. Robust benchmarks are vital as they provide standardized measures to objectively compare models, ensuring reliability and accuracy in financial text understanding and prediction. This systematic evaluation fosters transparency, reproducibility, and continuous improvement in LLM applications. Sharing code and methodologies promotes collaboration, driving innovation and practical implementation in real-world financial scenarios.

A notable work in this field is **FLUE** (Shah et al. 2022), which denotes Financial Language Understanding Evaluation, addressing the unique challenges posed by financial texts. It is a comprehensive suite of benchmarks designed to assess the performance of language models on various financial NLP tasks. FLUE consists of five tasks: 1) financial sentiment analysis using the FPB dataset, 2) news headline

EXHIBIT 6

Benchmarks of LLMs on Financial Applications

Name	Year	Task	Modality	Model	Language	Open Source
PIXIU Xie, Han, Zhang et al. (2023), Xie et al. (2024)	2023	Multiple financial NLP tasks; stock prediction	Text, Table, Time-series	FinMA	Chinese, English	Yes[a]
FLUE Shah et al. (2022)	2022	Multiple financial NLP tasks	Text	FLANG	English	Yes[b]
AlphaFin Li et al. (2024)	2024	Financial question answering; stock prediction	Text	Stock-Chain	Chinese, English	Yes[c]
Li, Chan et al. (2023)	2023	Multiple financial NLP tasks	Text	–	English	–
BizBench Koncel-Kedziorski et al. (2023)	2023	Multiple financial NLP tasks; program synthesis	Text, Table, Code	–	English	Yes[d]
DOCMATH-EVAL Zhao, Long et al. (2023)	2023	Numerical reasoning	Text, Table	–	English	Yes[e]
EconLogicQA Quan and Liu (2024)	2024	Financial question answering	Text	–	English	Yes[f]
FINANCEBENCH Islam et al. (2023)	2023	Financial question answering	Text	–	English	Yes[g]
Lakkaraju et al. (2023)	2023	Financial advisement	Text	–	English	–
MultiLing 2019 El-Haj (2019)	2019	Financial narrative summarization	Text	–	English	Yes[h]
R-Judge Yuan et al. (2024)	2024	Safety judgment; risk identification	Text	–	English	Yes[i]
BBT-Fin Lu et al. (2023)	2023	Multiple financial NLP tasks	Text	BBT-FinT5	Chinese	Yes[j]
CFBenchmark Lei et al. (2023)	2024	Multiple financial NLP tasks	Text	–	Chinese	Yes[k]
Hirano (2024)	2024	Multiple financial NLP tasks	Text	–	Japanese	Yes[l]
FLARE-ES Zhang, Xiang et al. (2024)	2024	Multiple financial NLP tasks	Text, Table, Time-series	FinMA-ES	Spanish, English	Yes[a]
FinEval Zhang, Cai et al. (2023)	2023	Financial domain knowledge	Text	–	Chinese	Yes[m]
ICE-PIXIU Hu et al. (2024)	2024	Multiple financial NLP tasks	Text, Table, Time-series	ICE-INTENT	Chinese, English	Yes[a]
SuperCLUE-Fin Xu, Zhu et al. (2024)	2024	Various financial tasks	Text	–	Chinese	Yes[n]

NOTES: Open-source links: [a]https://github.com/The-FinAI/PIXIU [b]https://salt-nlp.github.io/FLANG/ [c]https://github.com/AlphaFin-proj/AlphaFin [d]https://huggingface.co/datasets/kensho/BizBench [e]https://github.com/yale-nlp/DocMath-Eval [f]https://huggingface.co/datasets/yinzhu-quan/econlogicqa [g]https://github.com/patronus-ai/financebench [h]http://multiling.iit.demokritos.gr/pages/view/1754/multiling-2019 [i]https://github.com/Lordog/R-Judge [j]https://github.com/ssymmetry/BBT-FinCUGE-Application [k]https://github.com/TongjiFinLab/CFBenchmark [l]https://github.com/pfnet-research/japanese-lm-financial-benchmark [m]https://github.com/SUFE-AIFLM-Lab/FinEval [n]https://www.CLUEbenchmarks.com.

classification based on the gold news headline dataset, 3) named entity recognition with data from financial agreements, 4) structure boundary detection using the FinSBD dataset, and 5) question answering with data from the FiQA challenge. In addition, this article introduces FLANG-BERT and FLANG-ELECTRA, two models specifically trained on financial data using a novel pretraining methodology that incorporates financial keywords and phrases for better masking, as well as span boundary and in-filing objectives. These benchmarks cover a range of tasks critical for financial NLP, providing a robust platform to evaluate the effectiveness of financial language models.

PIXIU (Xie, Han, Zhang et al. 2023) represents a more recent development in the field, introducing a comprehensive framework that includes a financial LLM called FinMA, a large-scale multitask instruction dataset, and a holistic evaluation benchmark named FLARE (Financial Language Understanding and Prediction Evaluation Benchmark). PIXIU is characterized by its open resources, making all components, including the model, instruction-tuning data, and benchmarks, publicly available to promote transparency and further research. The instruction-tuning data in PIXIU covers various financial tasks and modalities, including text, tables, and time series data, ensuring comprehensive model training. The FLARE benchmark evaluates models on four financial NLP tasks (sentiment analysis, news headline classification, named entity recognition, and question answering) and one financial prediction task (stock movement prediction), covering nine datasets in total. This broad assessment allows for a thorough evaluation of a model's capabilities in handling diverse financial data, providing a more holistic benchmark compared to previous ones focused solely on NLP.

In addition, various other benchmarks have been developed to evaluate LLMs on a wide range of financial tasks. These benchmarks are closely related to the real-world applications that we discussed in the previous sections, including linguistic tasks, sentiment analysis, numerical reasoning, and comprehensive financial analysis. For example, Li, Chan et al. (2023) explore the effectiveness of LLMs in financial text analytics. **MultiLing 2019** (El-Haj 2019) and **BizBench** (Koncel-Kedziorski et al. 2023) evaluate models on their ability to summarize financial narratives and perform quantitative reasoning in business and finance contexts. For interpretable financial prediction, benchmarks such as **AlphaFin** (Li et al. 2024) and **FinanceBench** (Islam et al. 2023) assess models on stock trend prediction and financial question answering. Numerical reasoning capabilities are evaluated using benchmarks like **DocMath-Eval** (Zhao, Long et al. 2023), which test models on interpreting and calculating complex financial data from long documents. Comprehensive benchmarks like **R-Judge** (Yuan et al. 2024) and **EconLogicQA** (Quan and Liu 2024) focus on assessing risk awareness, safety in financial decision-making, and sequential reasoning within economic contexts. Xu, Zhou et al. (2024) present **FinTruthQA**, a dataset for evaluating the quality of financial disclosures in Chinese stock exchange Q&A platforms, focusing on NLP models to improve transparency and trust in financial markets. Together, these benchmarks provide a promising development for evaluating the diverse capabilities of LLMs in financial applications, ensuring models are tested across a broad spectrum of tasks.

Impact of language. Besides the aforementioned benchmarks, the language impact on the performance of financial LLMs has become another topic of interest and has been extensively explored. This research often focuses on creating benchmarks for specific languages or comparing model performance across different languages to understand their effectiveness in diverse linguistic contexts.

Several benchmarks have been developed to evaluate models on tasks such as sentiment analysis, named entity recognition, relation extraction, and financial news summarization in the Chinese financial domain. Benchmarks such as **BBT-Fin** (Lu et al. 2023) and **CFBenchmark** (Lei et al. 2023) are designed to provide comprehensive datasets and evaluation frameworks tailored to the linguistic and financial nuances of Chinese texts. Similarly, **FinEval** (Zhang, Cai et al. 2023) and **SuperCLUE-Fin** (Xu, Zhu et al. 2024) focus on a broader range of financial tasks, advancing Chinese financial NLP by addressing both theoretical knowledge and practical applications such as compliance, risk management, and investment analysis.

In the Japanese context, benchmarks such as the one developed by Hirano (2024) evaluate models on tasks like sentiment analysis, auditing tasks from the Japanese CPA (Certified Public Accountant) exam, and financial planner exam questions. This benchmark provides a robust framework to assess models' proficiency in Japanese financial texts.

Furthermore, several researchers explore bilingual capabilities to examine the performance of financial LLMs between different languages. Zhang, Xiang et al. (2024) focus on the comparison between Spanish and English, highlighting the challenges and effectiveness of models in processing and understanding financial texts across these languages. Hu et al. (2024) extend this comparison to Chinese and English, providing insights into the models' generalization and adaptation capabilities across diverse linguistic contexts.

These language-specific benchmarks and comparative studies are crucial for understanding the linguistic impact on financial LLMs. They ensure that models are capable of accurately processing and interpreting financial information in various major languages, thereby broadening their applicability and effectiveness in global financial markets.

CONCLUSION

This article provides a comprehensive overview of the application of LLMs in the financial domain, highlighting their capabilities in enhancing various financial tasks such as linguistic tasks, sentiment analysis, financial time series analysis, financial reasoning, and agent-based modeling. To support further research in this field, we have also compiled a useful collection of relevant datasets, code repositories, and benchmarks.

LLMs demonstrate remarkable potential in improving the efficiency and accuracy of financial processes through advanced contextual understanding and real-time analysis. Despite their promising capabilities, challenges such as data privacy, interpretability, and computational costs need to be addressed to ensure the responsible and effective deployment of LLMs in finance. As research continues to evolve, it is our hope that this review will encourage more exploration and discussion on the potential and limitations of LLMs, advancing their integration into the financial sector for more strategic investment management and efficient decision-making.

AUTHOR NOTE

Y. Kong and Y. Nie contributed equally to this article and are listed in random order. The remaining authors are listed in alphabetical order by surname. J. M. Mulvey (corresponding author) and S. Zohren are main advisors.

ACKNOWLEDGMENTS

Y. N. and H. V. Poor acknowledge financial support from Princeton Language and Intelligence at Princeton University. X.D. acknowledges support from the Oxford-Man Institute of Quantitative Finance.

REFERENCES

Abdaljalil, S., and H. Bouamor. 2021. "An Exploration of Automatic Text Summarization of Financial Reports." In *Proceedings of the Third Workshop on Financial Technology and Natural Language Processing,* pp. 1–7.

Aggarwal, S., S. Mehra, and P. Mitra. 2023. "Multi-Purpose NLP Chatbot: Design, Methodology & Conclusion." arXiv preprint arXiv:2310.08977.

Alias, M. S., M. H. Fuad, X. L. F. Hoong, and E. G. Y. Hin. 2023. "Financial Text Categorisation with FinBERT on Key Audit Matters." In *2023 IEEE Symposium on Computers & Informatics (ISCI),* pp. 63–69. IEEE.

Alvarado, J. C. S., K. Verspoor, and T. Baldwin. 2015. "Domain Adaption of Named Entity Recognition to Support Credit Risk Assessment." In *Proceedings of the Australasian Language Technology Association Workshop 2015*, pp. 84–90.

Anastasiou, D., and A. Katsafados. 2023. "Bank Deposits and Textual Sentiment: When an European Central Bank President's Speech Is Not Just a Speech." *The Manchester School* 91 (1): 55–87.

Aparicio, V., D. Gordon, S. G. Huayamares, and Y. Luo. 2024. "BioFinBERT: Finetuning Large Language Models (LLMs) to Analyze Sentiment of Press Releases and Financial Text around Inflection Points of Biotech Stocks." arXiv preprint arXiv:2401.11011.

Araci, D. 2019. "FinBERT: Financial Sentiment Analysis with Pre-Trained Language Models." arXiv preprint arXiv:1908.10063.

Arslan, Y., K. Allix, L. Veiber, C. Lothritz, T. F. Bissyandé, J. Klein, and A. Goujon. 2021. "A Comparison of Pre-Trained Language Models for Multi-Class Text Classification in the Financial Domain." In *Companion Proceedings of the Web Conference 2021*, pp. 260–268.

Arun, A., A. Dhiman, M. Soni, and Y. Hu. 2023. "Numerical Reasoning for Financial Reports." arXiv preprint arXiv:2312.14870.

Arvanitis, K., and N. Bassiliades. 2017. "Real-Time Investors' Sentiment Analysis from Newspaper Articles." In *Advances in Combining Intelligent Methods: Postproceedings of the 5th International Workshop CIMA-2015*, Vietri sul Mare, Italy, November 2015 (at ICTAI 2015), pp. 1–23. Berlin, Germany: Springer.

Au, W., A. Ait-Azzi, and J. Kang. 2021. "FinSBD-2021: The 3rd Shared Task on Structure Boundary Detection in Unstructured Text in the Financial Domain." In *Companion Proceedings of the Web Conference 2021*, pp. 276–279.

Avramelou, L., N. Passalis, G. Tsoumakas, and A. Tefas. 2023. "Domain-Specific Large Language Model Finetuning Using a Model Assistant for Financial Text Summarization." In *2023 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 381–386. IEEE.

Băroiu, A.-C., and Ş. Trăuşan-Matu. 2023. "How Capable Are State-of-the-Art Language Models to Cope with Sarcasm?" In *2023 24th International Conference on Control Systems and Computer Science (CSCS),* pp. 399–402. IEEE.

Beltagy, I., M. E. Peters, and A. Cohan. 2020. "Longformer: The Long-Document Transformer." arXiv preprint arXiv:2004.05150.

Berger, A., L. Hillebrand, D. Leonhard, T. Deußer, T. B. F. De Oliveira, T. Dilmaghani, M. Khaled, B. Kliem, R. Loitz, C. Bauckhage, et al. 2023. "Towards Automated Regulatory Compliance Verification in Financial Auditing with Large Language Models." In *2023 IEEE International Conference on Big Data (BigData)*, pp. 4626–4635. IEEE.

Bhatia, G., E. M. B. Nagoudi, H. Cavusoglu, and M. Abdul-Mageed. 2024. "FinTral: A Family of GPT-4 Level Multimodal Financial Large Language Models." arXiv preprint arXiv:2402.10986.

Bhattacharya, I., and A. Mickovic. 2024. “Accounting Fraud Detection Using Contextual Language Learning.” *International Journal of Accounting Information Systems* 53: 100682.

Biswas, S., A. Ghosh, S. Chakraborty, S. Roy, and R. Bose. 2020. “Scope of Sentiment Analysis on News Articles Regarding Stock Market and GDP in Struggling Economic Condition.” *International Journal* 8 (7): 3594–3609.

Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov. 2017. “Enriching Word Vectors with Subword Information.” *Transactions of the Association for Computational Linguistics* 5: 135–146.

Bordoloi, M., and S. K. Biswas. 2023. “Sentiment Analysis: A Survey on Design Framework, Applications and Future Scopes.” *Artificial Intelligence Review* 56 (March): 12505–12560.

Bosancic, T., Y. Nie, and J. M. Mulvey. 2024. “Regime-Aware Factor Allocation with Optimal Feature Selection.” Working paper, SSRN 4825234.

Burke, J., R. Hoitash, U. Hoitash, and S. X. Xiao. 2023. “Using a Large Language Model for Accounting Topic Classification.” Working paper, SSRN 4484489.

Cao, S., W. Jiang, B. Yang, and A. L. Zhang. 2023. “How to Talk When a Machine Is Listening: Corporate Disclosure in the Age of AI.” *The Review of Financial Studies* 36 (9): 3603–3642.

Cao, Y., Z. Chen, Q. Pei, F. Dimino, L. Ausiello, P. Kumar, K. Subbalakshmi, and P. M. Ndiaye. 2024. “RiskLabs: Predicting Financial Risk Using Large Language Model Based on Multi-Sources Data.” arXiv preprint arXiv:2404.07452.

Cao, Z., and Z. Feinstein. 2024. “Large Language Model in Financial Regulatory Interpretation.” arXiv preprint arXiv:2405.06808.

Caspi, I., S. S. Felber, and T. B. Gillis. 2023. “Generative AI and the Future of Financial Advice Regulation.” GenLaw Center. https://blog.genlaw.org/CameraReady/19.pdf.

Chai, Y., M. Chen, H. Wu, and S. Wang. 2023. “Fin-EMRC: An Efficient Machine Reading Comprehension Framework for Financial Entity-Relation Extraction.” *IEEE Access* (July 28).

Chandola, V., A. Banerjee, and V. Kumar. 2009. “Anomaly Detection: A Survey.” *ACM Computing Surveys (CSUR)* 41 (3): 1–58.

Chen, J., S. Yuan, R. Ye, B. P. Majumder, and K. Richardson. 2023. “Put Your Money Where Your Mouth Is: Evaluating Strategic Planning and Execution of LLM Agents in an Auction Arena.” arXiv preprint arXiv:2310.05746.

Chen, S., and F. Xing. 2023. “Understanding Emojis for Financial Sentiment Analysis.” Conference paper, *International Conference on Information Systems*, Hyderabad (December), 1284.

Chen, W., Y. Xu, Z. Zheng, Y. Zhou, J. E. Yang, and J. Bian. 2019. “Detecting ‘Pump & Dump Schemes’ on Cryptocurrency Market Using an Improved Apriori Algorithm.” In *2019 IEEE International Conference on Service-Oriented System Engineering (SOSE)*, pp. 293–295.

Chen, Y., Y. Gel, and H. V. Poor. 2022. “Time-Conditioned Dances with Simplicial Complexes: Zigzag Filtration Curve Based Supra-Hodge Convolution Networks for Time-Series Forecasting.” *Advances in Neural Information Processing Systems* 35: 8940–8953.

Chen, Y., B. T. Kelly, and D. Xiu. 2022. “Expected Returns and Large Language Models.” Working paper, SSRN 4416687.

Chen, Z., W. Chen, C. Smiley, S. Shah, I. Borova, D. Langdon, R. Moussa, M. Beane, T.-H. Huang, B. Routledge, et al. 2021. “FinQA: A Dataset of Numerical Reasoning Over Financial Data.” arXiv preprint arXiv:2109.00122.

Chen, Z., L. N. Zheng, C. Lu, J. Yuan, and D. Zhu. 2023. “ChatGPT Informed Graph Neural Network for Stock Movement Prediction.” arXiv preprint arXiv:2306.03763.

Cheng, Z., L. Wu, T. Lukasiewicz, E. Sallinger, and G. Gottlob. 2022. "Democratizing Financial Knowledge Graph Construction by Mining Massive Brokerage Research Reports." In *Workshop Proceedings of the EDBT/ICDT 2022 Joint Conference* (March 29–April 1, 2022), Edinburgh.

Chiong, R., Z. Fan, Z. Hu, M. T. Adam, B. Lutz, and D. Neumann. 2018. "A Sentiment Analysis-Based Machine Learning Approach for Financial Market Prediction via News Disclosures." In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pp. 278–279.

Choi, G.-Y., and A. G. Kim. 2024. "Firm-Level Tax Audits: A Generative AI-Based Measurement." Chicago Booth Research Paper No. 23-23. arXiv:2406.11903.

Chuang, C., and Y. Yang. 2022. "Buy Tesla, Sell Ford: Assessing Implicit Stock Market Preference in Pre-Trained Language Models." In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 100–105.

Consoli, S., L. Barbaglia, and S. Manzan. 2022. "Fine-Grained, Aspect-Based Sentiment Analysis on Economic and Financial Lexicon." *Knowledge-Based Systems* 247: 108781.

Cook, T. R., S. Kazinnik, A. L. Hansen, and P. McAdam. 2023. "Evaluating Local Language Models: An Application to Financial Earnings Calls." Working paper, SSRN 4627143.

Covas, E. 2023. "Named Entity Recognition Using GPT for Identifying Comparable Companies." arXiv preprint arXiv:2307.07420.

Crépey, S., N. Lehdili, N. Madhar, and M. Thomas. 2022. "Anomaly Detection in Financial Time Series by Principal Component Analysis and Neural Networks." *Algorithms Special Issue Machine Learning and Deep Learning Applications for Anomaly and Fault Detection* 15 (10): 385.

Curti, F., and S. Kazinnik. 2023. "Let's Face It: Quantifying the Impact of Nonverbal Communication in FOMC Press Conferences." *Journal of Monetary Economics* 139: 110–126.

Darban, Z. Z., G. I. Webb, S. Pan, C. C. Aggarwal, and M. Salehi. 2022. "Deep Learning for Time Series Anomaly Detection: A Survey." arXiv preprint arXiv:2211.05244.

de Zarzà, I., J. de Curtò, G. Roig, and C. T. Calafate. 2023. "Optimized Financial Planning: Integrating Individual and Cooperative Budgeting Models with LLM Recommendations." *AI* 5 (1): 91–114.

Deng, X., V. Bashlovkina, F. Han, S. Baumgartner, and M. Bendersky. 2023. "What Do LLMs Know about Financial Markets? A Case Study on Reddit Market Sentiment Analysis." In *Companion Proceedings of the ACM Web Conference 2023*, pp. 107–110.

Deußer, T., D. Leonhard, L. Hillebrand, A. Berger, M. Khaled, S. Heiden, T. Dilmaghani, B. Kliem, R. Loitz, C. Bauckhage, et al. 2023. "Uncovering Inconsistencies and Contradictions in Financial Reports Using Large Language Models." In *2023 IEEE International Conference on Big Data (BigData)*, pp. 2814–2822. IEEE.

Dickinson, B., and W. Hu. 2015. "Sentiment Analysis of Investor Opinions on Twitter." *Social Networking* 4 (03): 62.

Ding, B., C. Qin, R. Zhao, T. Luo, X. Li, G. Chen, W. Xia, J. Hu, A. T. Luu, and S. Joty. 2024. "Data Augmentation Using LLMs: Data Perspectives, Learning Paradigms and Challenges." arXiv preprint arXiv:2403.02990.

Dong, M. M., T. C. Stratopoulos, and V. X. Wang. 2024. "A Scoping Review of ChatGPT Research in Accounting and Finance." *International Journal of Accounting Information Systems*, December. https://doi.org/10.1016/j.accinf.2024.100715, Available at SSRN: https://ssrn.com/abstract=4680203 or http://dx.doi.org/10.2139/ssrn.4680203.

Ebrahimi, M., A. H. Yazdavar, and A. Sheth. 2017. "Challenges of Sentiment Analysis for Dynamic Events." *IEEE Intelligent Systems* 32 (5): 70–75.

Eddy, S. R. 1996. "Hidden Markov Models." *Current Opinion in Structural Biology* 6 (3): 361–365.

Ehrmann, M., A. Hamdi, E. L. Pontes, M. Romanello, and A. Doucet. 2023. "Named Entity Recognition and Classification in Historical Documents: A Survey." *ACM Computing Surveys* 56 (2): 1–47.

El-Haj, M. 2019. "MultiLing 2019: Financial Narrative Summarisation." In *Proceedings of the Workshop MultiLing 2019: Summarization Across Languages, Genres and Sources*, pp. 6–10.

Epstein, J. M. 1999. "Agent-Based Computational Models and Generative Social Science." *Complexity* 4 (5): 41–60.

Fabozzi, F. A. 2024. "A New Framework for the Application of Generative Language Models for Portfolio Construction." PhD thesis, Stevens Institute of Technology.

Farmer, J. D., and D. Foley. 2009. "The Economy Needs Agent-Based Modelling." *Nature* 460 (7256): 685–686.

Fatouros, G., K. Metaxas, J. Soldatos, and D. Kyriazis. 2024. "Can Large Language Models Beat Wall Street? Unveiling the Potential of AI in Stock Selection." arXiv preprint arXiv:2401.03737.

Fatouros, G., J. Soldatos, K. Kouroumali, G. Makridis, and D. Kyriazis. 2023. "Transforming Sentiment Analysis in the Financial Domain with ChatGPT." *Machine Learning with Applications* 14: 100508.

Fava, D. 2023. "The Future of Tax Planning: Leveraging Generative AI in High-Net-Worth Contexts." *Journal of Financial Planning* 36 (10): 54–57.

Feng, D., Y. Dai, J. Huang, Y. Zhang, Q. Xie, W. Han, A. Lopez-Lira, and H. Wang. 2023. "Empowering Many, Biasing a Few: Generalist Credit Scoring through Large Language Models." arXiv preprint arXiv:2310.00566.

Foroutan, N., A. Romanou, S. Massonnet, R. Lebret, and K. Aberer. 2022. "Multilingual Text Summarization on Financial Documents." In *Proceedings of the 4th Financial Narrative Processing Workshop@ LREC2022*, pp. 53–58.

Ghosh, S., S. Umrao, C.-C. Chen, and S. K. Naskar. 2023. "The Mask One at a Time Framework for Detecting the Relationship between Financial Entities." In *Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation*, pp. 40–43.

Gössi, S., Z. Chen, W. Kim, B. Bermeitinger, and S. Handschuh. 2023. "FinBERT-FOMC: Fine-Tuned FinBERT Model with Sentiment Focus Method for Enhancing Sentiment Analysis of FOMC Minutes." In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pp. 357–364.

Gruver, N., M. Finzi, S. Qiu, and A. G. Wilson. 2024. "Large Language Models Are Zero-Shot Time Series Forecasters." *Advances in Neural Information Processing Systems* 36: 19622–19635.

Guo, T., X. Chen, Y. Wang, R. Chang, S. Pei, N. V. Chawla, O. Wiest, and X. Zhang. 2024. "Large Language Model Based Multi-Agents: A Survey of Progress and Challenges." arXiv preprint arXiv:2402.01680.

Gupta, H., S. Verma, S. Mashetty, and S. Mishra. 2021. "Context-NER: Contextual Phrase Generation at Scale." arXiv preprint arXiv:2109.08079.

Gupta, U. 2023. "GPT-InvestAR: Enhancing Stock Investment Strategies through Annual Report Analysis with Large Language Models." arXiv preprint arXiv:2309.03079.

Hadi, M. U., R. Qureshi, A. Shah, M. Irfan, A. Zafar, M. B. Shaikh, N. Akhtar, J. Wu, S. Mirjalili, et al. 2024. "Large Language Models: A Comprehensive Survey of Its Applications, Challenges, Limitations, and Future Prospects." *TechRxiv* (September 5). https://www.techrxiv.org/users/618307/articles/682263-large-language-models-a-comprehensive-survey-of-its-applications-challenges-limitations-and-future-prospects.

Hearst, M., S. Dumais, E. Osuna, J. Platt, and B. Scholkopf. 1998. "Support Vector Machines." *IEEE Intelligent Systems and Their Applications* 13 (4): 18–28.

Hillebrand, L., A. Berger, T. Deußer, T. Dilmaghani, M. Khaled, B. Kliem, R. Loitz, M. Pielka, D. Leonhard, C. Bauckhage, et al. 2023. "Improving Zero-Shot Text Matching for Financial Auditing with Large Language Models." In *Proceedings of the ACM Symposium on Document Engineering 2023*, pp. 1–4.

Hillebrand, L., T. Deußer, T. Dilmaghani, B. Kliem, R. Loitz, C. Bauckhage, and R. Sifa. 2022. "KPI-BERT: A Joint Named Entity Recognition and Relation Extraction Model for Financial Reports." In *2022 26th International Conference on Pattern Recognition (ICPR)*, pp. 606–612. IEEE.

Hirano, M. 2024. "Construction of a Japanese Financial Benchmark for Large Language Models." arXiv preprint arXiv:2403.15062.

Horton, J. J. 2023. "Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?" Technical report, National Bureau of Economic Research.

Hu, G., K. Qin, C. Yuan, M. Peng, A. Lopez-Lira, B. Wang, S. Ananiadou, W. Yu, J. Huang, and Q. Xie. 2024. "No Language Is an Island: Unifying Chinese and English in Financial Large Language Models, Instruction Data, and Benchmarks." arXiv preprint arXiv:2403.06249.

Hua, W., X. Yang, Z. Li, C. Wei, and Y. Zhang. 2024. "TrustAgent: Towards Safe and Trustworthy LLM-Based Agents through Agent Constitution." arXiv preprint arXiv:2402.01586.

Huang, Z., C. Che, H. Zheng, and C. Li. 2024. "Research on Generative Artificial Intelligence for Virtual Financial Robo-Advisor." *Academic Journal of Science and Technology* 10 (1): 74–80.

Islam, P., A. Kannappan, D. Kiela, R. Qian, N. Scherrer, and B. Vidgen. 2023. "FinanceBench: A New Benchmark for Financial Question Answering." arXiv preprint arXiv:2311.11944.

Jeong, C. 2024. "Fine-Tuning and Utilization Methods of Domain-Specific LLMs." arXiv preprint arXiv:2401.02981.

Jiang, X., C. Xu, Y. Shen, X. Sun, L. Tang, S. Wang, Z. Chen, Y. Wang, and J. Guo. 2023. "On the Evolution of Knowledge Graphs: A Survey and Perspective." arXiv preprint arXiv:2310.04835.

Jiang, Y., Z. Pan, X. Zhang, S. Garg, A. Schneider, Y. Nevmyvaka, and D. Song. 2024. "Empowering Time Series Analysis with Large Language Models: A Survey." arXiv preprint arXiv:2402.03182.

Jin, M., S. Wang, L. Ma, Z. Chu, J. Y. Zhang, X. Shi, P.-Y. Chen, Y. Liang, Y.-F. Li, S. Pan, et al. 2023. "Time-LLM: Time Series Forecasting by Reprogramming Large Language Models." arXiv preprint arXiv:2310.01728.

Jin, M., Q. Wen, Y. Liang, C. Zhang, S. Xue, X. Wang, J. Zhang, Y. Wang, H. Chen, X. Li, et al. 2023. "Large Models for Time Series and Spatio-Temporal Data: A Survey and Outlook." arXiv preprint arXiv:2310.10196.

Jin, M., Y. Zhang, W. Chen, K. Zhang, Y. Liang, B. Yang, J. Wang, S. Pan, and Q. Wen. 2024. "Position Paper: What Can Large Language Models Tell Us about Time Series Analysis." arXiv preprint arXiv:2402.02713.

Kalra, S., and J. S. Prasad. 2019. "Efficacy of News Sentiment for Stock Market Prediction." In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, pp. 491–496. IEEE.

Kanelis, D., and P. L. Siklos. 2024. "The ECB Press Conference Statement: Deriving a New Sentiment Indicator for the Euro Area." *International Journal of Finance & Economics*.

Kaur, S., C. Smiley, A. Gupta, J. Sain, D. Wang, S. Siddagangappa, T. Aguda, and S. Shah. 2023. "REFinD: Relation Extraction Financial Dataset." In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3054–3063.

Khanna, U., S. Ghodratnama, A. Beheshti, et al. 2022. "Transformer-Based Models for Long Document Summarisation in Financial Domain." In *Proceedings of the 4th Financial Narrative Processing Workshop@ LREC2022*, pp. 73–78.

Kim, A., M. Muhn, and V. V. Nikolaev. 2023. “Bloated Disclosures: Can ChatGPT Help Investors Process Information?” Chicago Booth Research Paper No. 23-07.

——. 2024. “Financial Statement Analysis with Large Language Models.” Chicago Booth Research Paper No. 2024-65; Fama-Miller Working Paper.

Kim, S., S. Kim, Y. Kim, J. Park, S. Kim, M. Kim, C. H. Sung, J. Hong, and Y. Lee. 2023. “LLMs Analyzing the Analysts: Do BERT and GPT Extract More Value from Financial Analyst Reports?” In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pp. 383–391.

Kim, W., J. F. Spörer, and S. Handschuh. 2023. “Analyzing FOMC Minutes: Accuracy and Constraints of Language Models.” arXiv preprint arXiv:2304.10164.

Kirange, D., and R. R. Deshmukh. 2016. “Sentiment Analysis of News Headlines for Stock Price Prediction.” *Composoft, An International Journal of Advanced Computer Technology* 5 (3): 2080–2084.

Klejdysz, J., and R. L. Lumsdaine. 2023. “Shifts in ECB Communication: A Textual Analysis of the Press Conference.” *International Journal of Central Banking* 19 (2): 473–542.

Ko, H., and J. Lee. 2024. “Can ChatGPT Improve Investment Decisions? From a Portfolio Management Perspective.” *Finance Research Letters* 64: 105433.

Koa, K. J., Y. Ma, R. Ng, and T.-S. Chua. 2024. “Learning to Generate Explainable Stock Predictions Using Self-Reflective Large Language Models.” arXiv preprint arXiv:2402.03659.

Koncel-Kedziorski, R., M. Krumdick, V. Lai, V. Reddy, C. Lovering, and C. Tanner. 2023. “BizBench: A Quantitative Reasoning Benchmark for Business and Finance.” arXiv preprint arXiv:2311.06602.

Kong, Y., Z. Wang, Y. Nie, T. Zhou, S. Zohren, Y. Liang, P. Sun, and Q. Wen. 2024. “Unlocking the Power of LSTM for Long Term Time Series Forecasting.” arXiv preprint arXiv:2408.10006.

La Quatra, M., and L. Cagliero. 2020. “End-to-End Training for Financial Report Summarization.” In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pp. 118–123.

Lakkaraju, K., S. E. Jones, S. K. R. Vuruma, V. Pallagani, B. C. Muppasani, and B. Srivastava. 2023. “LLMs for Financial Advisement: A Fairness and Efficacy Study in Personal Decision Making.” In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pp. 100–107.

Lakkaraju, K., S. K. R. Vuruma, V. Pallagani, B. Muppasani, and B. Srivastava. 2023. “Can LLMs Be Good Financial Advisors? An Initial Study in Personal Decision Making for Optimized Outcomes.” arXiv preprint arXiv:2307.07422.

Le, Q., and T. Mikolov. 2014. “Distributed Representations of Sentences and Documents.” In *International Conference on Machine Learning*, pp. 1188–1196. PMLR.

Lee, J., and M. Kim. 2023. “ESG Information Extraction with Cross-Sectoral and Multi-Source Adaptation Based on Domain-Tuned Language Models.” *Expert Systems with Applications* 221: 119726.

Lee, J., N. Stevens, S. C. Han, and M. Song. 2024. “A Survey of Large Language Models in Finance (FinLLMs).” arXiv preprint arXiv:2402.02315.

Lei, Y., J. Li, M. Jiang, J. Hu, D. Cheng, Z. Ding, and C. Jiang. 2023. “CFBenchmark: Chinese Financial Assistant Benchmark for Large Language Model.” arXiv preprint arXiv:2311.05812.

Leippold, M. 2023. “Sentiment Spin: Attacking Financial Sentiment with GPT-3.” *Finance Research Letters* 55: 103957.

Li, H., H. H. Gao, C. Wu, and M. A. Vasarhelyi. 2023. “Extracting Financial Data from Unstructured Sources: Leveraging Large Language Models.” Working paper, SSRN.

Li, J., A. Sun, J. Han, and C. Li. 2020. “A Survey on Deep Learning for Named Entity Recognition.” *IEEE Transactions on Knowledge and Data Engineering* 34 (1): 50–70.

Li, N., C. Gao, Y. Li, and Q. Liao. 2023. "EconAgent: Large Language Model-Empowered Agents for Simulating Macroeconomic Activities." arXiv preprint arXiv:2310.10436.

Li, Q., and Q. Zhang. 2021. "A Unified Model for Financial Event Classification, Detection and Summarization." In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pp. 4668–4674.

Li, X., S. Chan, X. Zhu, Y. Pei, Z. Ma, X. Liu, and S. Shah. 2023. "Are ChatGPT and GPT-4 General-Purpose Solvers for Financial Text Analytics? A Study on Several Typical Tasks." In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 408–422.

Li, X., Z. Li, C. Shi, Y. Xu, Q. Du, M. Tan, J. Huang, and W. Lin. 2024. "AlphaFin: Benchmarking Financial Analysis with Retrieval-Augmented Stock-Chain Framework." arXiv preprint arXiv:2403.12582.

Li, X. V. 2023. "FinDKG: Dynamic Knowledge Graph with Large Language Models for Global Finance." Working paper, SSRN 4608445.

Li, Y., S. Wang, H. Ding, and H. Chen. 2023. "Large Language Models in Finance: A Survey." In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pp. 374–382.

Li, Y., Y. Yu, H. Li, Z. Chen, and K. Khashanah. 2023. "TradingGPT: Multi-Agent System with Layered Memory and Distinct Characters for Enhanced Financial Trading Performance." arXiv preprint arXiv:2309.03736.

Liang, Y., K. Tan, T. Xie, W. Tao, S. Wang, Y. Lan, and W. Qian. 2024. "Aligning Large Language Models to a Domain-Specific Graph Database." arXiv preprint arXiv:2402.16567.

Liang, Y., H. Wen, Y. Nie, Y. Jiang, M. Jin, D. Song, S. Pan, and Q. Wen. 2024. "Foundation Models for Time Series Analysis: A Tutorial and Survey." arXiv preprint arXiv:2403.14735.

Lim, B., and S. Zohren. 2021. "Time-Series Forecasting with Deep Learning: A Survey." *Philosophical Transactions of the Royal Society A* 379 (2194): 20200209.

Lipton, Z. C., J. Berkowitz, and C. Elkan. 2015. "A Critical Review of Recurrent Neural Networks for Sequence Learning." arXiv preprint arXiv:1506.00019.

Lopez-Lira, A., and Y. Tang. 2023. "Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models." arXiv preprint arXiv:2304.07619.

Loughran, T., and B. McDonald. 2011. "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks." *The Journal of Finance* 66 (1): 35–65.

Loukas, L., M. Fergadiotis, I. Chalkidis, E. Spyropoulou, P. Malakasiotis, I. Androutsopoulos, and G. Paliouras. 2022. "FiNER: Financial Numeric Entity Recognition for XBRL Tagging." arXiv preprint arXiv:2203.06482.

Loukas, L., I. Stogiannidis, O. Diamantopoulos, P. Malakasiotis, and S. Vassos. 2023. "Making LLMs Worth Every Penny: Resource-Limited Text Classification in Banking." In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pp. 392–400.

Lu, D., H. Wu, J. Liang, Y. Xu, Q. He, Y. Geng, M. Han, Y. Xin, and Y. Xiao. 2023. "BBT-Fin: Comprehensive Construction of Chinese Financial Domain Pre-Trained Language Model, Corpus and Benchmark." arXiv preprint arXiv:2302.09432.

Lu, F., L. Huang, and S. Li. 2023. "ChatGPT, Generative AI, and Investment Advisory." Working paper, SSRN.

Ludwig, E. T., and C. R. Bennetts. 2023. "Streamlining Financial Planning with ChatGPT: A Collaborative Approach between Technology and Human Expertise." *Journal of Financial Planning* 36 (6).

Luo, W., and D. Gong. 2024. "Pre-Trained Large Language Models for Financial Sentiment Analysis." arXiv preprint arXiv:2401.05215.

Ma, Q., X. Xue, D. Zhou, X. Yu, D. Liu, X. Zhang, Z. Zhao, Y. Shen, P. Ji, J. Li, et al. 2024. "Computational Experiments Meet Large Language Model Based Agents: A Survey and Perspective." arXiv preprint arXiv:2402.00262.

Mackie, I., and J. Dalton. 2022. "Query-Specific Knowledge Graphs for Complex Finance Topics." arXiv preprint arXiv:2211.04142.

Maia, M., S. Handschuh, A. Freitas, B. Davis, R. McDermott, M. Zarrouk, and A. Balahur. 2018. "WWW'18 Open Challenge: Financial Opinion Mining and Question Answering." In *Companion Proceedings of the Web Conference 2018*, pp. 1941–1942.

Malo, P., A. Sinha, P. Korhonen, J. Wallenius, and P. Takala. 2014. "Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts." *Journal of the Association for Information Science and Technology* 65 (4): 782–796.

Mehra, S., R. Louka, and Y. Zhang. 2022. "ESGBERT: Language Model to Help with Classification Tasks Related to Companies Environmental, Social, and Governance Practices." arXiv preprint arXiv:2203.16788.

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." arXiv preprint arXiv:1301.3781.

Mishev, K., A. Gjorgjevikj, I. Vodenska, L. T. Chitkushev, and D. Trajanov. 2020. "Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers." *IEEE Access* 8: 131662–131682.

Mishra, S. 2023. "ESG Impact Type Classification: Leveraging Strategic Prompt Engineering and LLM Fine-Tuning." In *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing*, pp. 72–78.

Mody, A., and M. Nedeljkovic. 2024. "Central Bank Policies and Financial Markets: Lessons from the Euro Crisis." *Journal of Banking & Finance* 158: 107033.

Mukherjee, R., A. Bohra, A. Banerjee, S. Sharma, M. Hegde, A. Shaikh, S. Shrivastava, K. Dasgupta, N. Ganguly, S. Ghosh, et al. 2022. "ECTSum: A New Benchmark Dataset for Bullet Point Summarization of Long Earnings Call Transcripts." arXiv preprint arXiv:2210.12467.

Mulvey, J. M., J. Gu, M. Holen, and Y. Nie. 2022. "Applications of Machine Learning in Wealth Management." *Journal of Investment Consulting* 21 (1): 66–82.

Mumtaz, U., and S. Mumtaz. 2023. "Potential of ChatGPT in Predicting Stock Market Trends Based on Twitter Sentiment Analysis." arXiv preprint arXiv:2311.06273.

Nadeau, D., and S. Sekine. 2007. "A Survey of Named Entity Recognition and Classification." *Lingvisticae Investigationes* 30 (1): 3–26.

Nagy, P., S. Frey, S. Sapora, K. Li, A. Calinescu, S. Zohren, and J. Foerster. 2023. "Generative AI for End-to-End Limit Order Book Modelling: A Token-Level Autoregressive Generative Model of Message Flow Using a Deep State Space Network." In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pp. 91–99.

Nguyen, S. T., and T. Tulabandhula. 2023. "Generative AI for Business Strategy: Using Foundation Models to Create Business Strategy Tools." arXiv preprint arXiv:2308.14182.

Ni, X., P. Li, and H. Li. 2023. "Unified Text Structuralization with Instruction-Tuned Language Models." arXiv preprint arXiv:2303.14956.

Nia, Z. M., A. Ahmadi, N. L. Bragazzi, W. A. Woldegerima, B. Mellado, J. Wu, J. Orbinski, A. Asgary, and J. D. Kong. 2022. "A Cross-Country Analysis of Macroeconomic Responses to COVID-19 Pandemic Using Twitter Sentiments." *PLOS ONE* 17 (8): e0272208.

Nie, Y., N. H. Nguyen, P. Sinthong, and J. Kalagnanam. 2022. "A Time Series Is Worth 64 Words: Long-Term Forecasting with Transformers." arXiv preprint arXiv:2211.14730.

Niszczota, P., and S. Abbas. 2023. "GPT Has Become Financially Literate: Insights from Financial Literacy Tests of GPT and a Preliminary Test of How People Use It as a Source of Advice." *Finance Research Letters* 58: 104333.

Noguer i Alonso, M., and H. Dupouy. 2024. "Evaluating LLMs in Financial Tasks—Code Generation in Trading Strategies." Working paper, SSRN 4752797 (March 8).

Nugroho, K. S., A. Y. Sukmadewa, and N. Yudistira. 2021. "Large-Scale News Classification Using BERT Language Model: Spark NLP Approach." In *Proceedings of the 6th International Conference on Sustainable Information Engineering and Technology*, pp. 240–246.

Ok, H. 2023. "FinTree: Financial Dataset Pretrain Transformer Encoder for Relation Extraction." arXiv preprint arXiv:2307.13900.

Ouyang, K., Y. Liu, S. Li, R. Bao, K. Harimoto, and X. Sun. 2024. "Modal-Adaptive Knowledge-Enhanced Graph-Based Financial Prediction from Monetary Policy Conference Calls with LLM." arXiv preprint arXiv:2403.16055.

Pakhale, K. 2023. "Comprehensive Overview of Named Entity Recognition: Models, Domain-Specific Applications and Challenges." arXiv preprint arXiv:2309.14084.

Pan, J. Z., S. Razniewski, J.-C. Kalo, S. Singhania, J. Chen, S. Dietze, H. Jabeen, J. Omeliyanenko, W. Zhang, M. Lissandrini, et al. 2023. "Large Language Models and Knowledge Graphs: Opportunities and Challenges." arXiv preprint arXiv:2308.06374.

Pan, S., L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu. 2024. "Unifying Large Language Models and Knowledge Graphs: A Roadmap." *IEEE Transactions on Knowledge and Data Engineering* 36 (7): 3580–3599.

Pan, Z., Y. Jiang, S. Garg, A. Schneider, Y. Nevmyvaka, and D. Song. 2024. "S^2IP-LLM: Semantic Space Informed Prompt Learning with LLM for Time Series Forecasting." arXiv preprint arXiv:2403.05798.

Pan, Z., Y. Jiang, D. Song, S. Garg, K. Rasul, A. Schneider, and Y. Nevmyvaka. 2024. "Structural Knowledge Informed Continual Multivariate Time Series Forecasting." arXiv preprint arXiv:2402.12722.

Park, T. 2024. "Enhancing Anomaly Detection in Financial Markets with an LLM-Based Multi-Agent Framework." arXiv preprint arXiv:2403.19735.

Pennebaker, J. W., M. E. Francis, and R. J. Booth. 2001. "Linguistic Inquiry and Word Count: LIWC 2001." Mahwah: Lawrence Erlbaum Associates 71.

Pennington, J., R. Socher, and C. D. Manning. 2014. "GloVe: Global Vectors for Word Representation." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543.

Phogat, K. S., C. Harsha, S. Dasaratha, S. Ramakrishna, and S. A. Puranam. 2023. "Zero-Shot Question Answering over Financial Documents Using Large Language Models." arXiv preprint arXiv:2311.14722.

Quan, Y., and Z. Liu. 2024. "EconLogicQA: A Question-Answering Benchmark for Evaluating Large Language Models in Economic Sequential Reasoning." arXiv preprint arXiv:2405.07938.

Quinlan, J. R. 1986. "Induction of Decision Trees." *Machine Learning* 1 (1): 81–106.

Raiaan, M. A. K., M. S. H. Mukta, K. Fatema, N. M. Fahad, S. Sakib, M. M. J. Mim, J. Ahmad, M. E. Ali, and S. Azam. 2024. "A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges." *IEEE Access*.

Rajpoot, P. K., and A. Parikh. 2023. "GPT-FinRE: In-Context Learning for Financial Relation Extraction Using Large Language Models." arXiv preprint arXiv:2306.17519.

Ramyadevi, R., and G. Sasidharan. 2024. “Cogniwealth: Revolutionizing Finance, Empowering Investors, and Shaping the Future of Wealth Management.” In *2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT)*, Vol. 5, pp. 378–381. IEEE.

Renault, T. 2020. “Sentiment Analysis and Machine Learning in Finance: A Comparison of Methods and Models on One Million Messages.” *Digital Finance* 2 (1): 1–13.

Rizinski, M., A. Jankov, V. Sankaradas, E. Pinsky, I. Mishkovski, and D. Trajanov. 2024. “Comparative Analysis of NLP-Based Models for Company Classification.” *Information* 15 (2): 77.

Rodriguez Inserte, P., M. Nakhlé, R. Qader, G. Caillaut, and J. Liu. 2024. “Large Language Model Adaptation for Financial Sentiment Analysis.” arXiv preprint arXiv:2401.14777.

Rosa, C. 2013. “The Financial Market Effect of FOMC Minutes.” *Economic Policy Review* 19 (2): 67–81.

Sarmah, B., D. Mehta, S. Pasquali, and T. Zhu. 2023. “Towards Reducing Hallucination in Extracting Information from Financial Reports Using Large Language Models.” In *Proceedings of the Third International Conference on AI-ML Systems*, pp. 1–5.

Sarzynska-Wawer, J., A. Wawer, A. Pawlak, J. Szymanowska, I. Stefaniak, M. Jarkiewicz, and L. Okruszek. 2021. “Detecting Formal Thought Disorder by Deep Contextualized Word Representations.” *Psychiatry Research* 304: 114135.

Shah, A., S. Paturi, and S. Chava. 2023. “Trillion Dollar Words: A New Financial Dataset, Task & Market Analysis.” arXiv preprint arXiv:2305.07972.

Shah, A., R. Vithani, A. Gullapalli, and S. Chava. 2023. “FiNER: Financial Named Entity Recognition Dataset and Weak-Supervision Model.” arXiv preprint arXiv:2302.11157.

Shah, R. S., K. Chawla, D. Eidnani, A. Shah, W. Du, S. Chava, N. Raman, C. Smiley, J. Chen, and D. Yang. 2022. “When FLUE Meets FLANG: Benchmarks and Large Pre-Trained Language Model for Financial Domain.” arXiv preprint arXiv:2211.00083.

Shapiro, A. H., M. Sudhof, and D. J. Wilson. 2022. “Measuring News Sentiment.” *Journal of Econometrics* 228 (2): 221–243.

Sharma, A., S. Rao, C. Brockett, A. Malhotra, N. Jojic, and W. B. Dolan. 2024. “Investigating Agency of LLMs in Human-AI Collaboration Tasks.” In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1968–1987.

Sharma, S., T. Nayak, A. Bose, A. K. Meena, K. Dasgupta, N. Ganguly, and P. Goyal. 2022. “FinRED: A Dataset for Relation Extraction in Financial Domain.” In *Companion Proceedings of the Web Conference 2022*, pp. 595–597.

Shukla, N. K., R. Katikeri, M. Raja, G. Sivam, S. Yadav, A. Vaid, and S. Prabhakararao. 2023. “Generative AI Approach to Distributed Summarization of Financial Narratives.” In *2023 IEEE International Conference on Big Data (BigData)*, pp. 2872–2876. IEEE.

Shukla, N., A. Vaid, R. Katikeri, S. Keeriyadath, and M. Raja. 2022. “DiMSum: Distributed and Multilingual Summarization of Financial Narratives.” In *Proceedings of the 4th Financial Narrative Processing Workshop@ LREC2022*, pp. 65–72.

Sims, C. A. 1980. “Macroeconomics and Reality.” *Econometrica*: 1–48.

Smales, L. A., and N. Apergis. 2017. “Does More Complex Language in FOMC Decisions Impact Financial Markets?” *Journal of International Financial Markets, Institutions and Money* 51: 171–189.

Smets, F., and R. Wouters. 2003. “An Estimated Dynamic Stochastic General Equilibrium Model of the Euro Area.” *Journal of the European Economic Association* 1 (5): 1123–1175.

Sohangir, S., N. Petty, and D. Wang. 2018. “Financial Sentiment Lexicon Analysis.” In *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, pp. 286–289. IEEE.

Sohangir, S., D. Wang, A. Pomeranets, and T. M. Khoshgoftaar. 2018. "Big Data: Deep Learning for Financial Sentiment Analysis." *Journal of Big Data* 5 (1): 1–25.

Srivastava, P., M. Malik, and T. Ganu. 2024. "Assessing LLMs' Mathematical Reasoning in Financial Document Question Answering." arXiv preprint arXiv:2402.11194.

Staudemeyer, R. C., and E. R. Morris. 2019. "Understanding LSTM–A Tutorial into Long Short-Term Memory Recurrent Neural Networks." arXiv preprint arXiv:1909.09586.

Steinert, R., and S. Altmann. 2023. "Linking Microblogging Sentiments to Stock Price Movement: An Application of GPT-4." arXiv preprint arXiv:2308.16771.

Stone, P. J., D. C. Dunphy, and M. S. Smith. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. Cambridge, MA: MIT Press.

Su, D., J. D. Lee, J. M. Mulvey, and H. V. Poor. 2022. "Competitive Multi-Agent Reinforcement Learning with Self-Supervised Representation." In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4098–4102. IEEE.

Su, D.-J., J. M. Mulvey, and H. V. Poor. 2022. "Improving Portfolio Performance Via Natural Language Processing Methods." *The Journal of Financial Data Science* 4 (2): 37–49.

Suzuki, M., H. Sakaji, M. Hirano, and K. Izumi. 2023. "Constructing and Analyzing Domain-Specific Language Model for Financial Text Mining." *Information Processing & Management* 60 (2): 103194.

Swaileh, W., T. Paquet, S. Adam, and A. Rojas Camacho. 2020. "A Named Entity Extraction System for Historical Financial Data." In *Document Analysis Systems*: *14th IAPR International Workshop, DAS 2020,* Wuhan, China, July 26–29, 2020, *Proceedings* 14, pp. 324–340. Berlin, Germany: Springer.

Taboada, M., J. Brooke, M. Tofiloski, K. Voll, and M. Stede. 2011. "Lexicon-Based Methods for Sentiment Analysis." *Computational Linguistics* 37 (2): 267–307.

Tan, K. L., C. P. Lee, and K. M. Lim. 2023. "A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research." *Applied Sciences* 13 (7): 4550.

Tesfatsion, L. 2006. "Agent-Based Computational Economics: A Constructive Approach to Economic Theory." *Handbook of Computational Economics* 2: 831–880.

Tian, C., Y. Zhao, and L. Ren. 2019. "A Chinese Event Relation Extraction Model Based on BERT." In *2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pp. 271–276. IEEE.

Touvron, H., T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. 2023. "LLaMA: Open and Efficient Foundation Language Models." arXiv preprint arXiv:2302.13971.

Trajanoska, M., R. Stojanov, and D. Trajanov. 2023. "Enhancing Knowledge Graph Construction Using Large Language Models." arXiv preprint arXiv:2305.04676.

Tsao, W.-K., and E. Chang. 2023. "Multi-Agent Reasoning with Large Language Models for Effective Corporate Planning." In *The 10th International Conference on Computational Science and Computational Intelligence* (October), Las Vegas.

Valencia, F., A. Gómez-Espinosa, and B. Valdés-Aguirre. 2019. "Price Movement Prediction of Cryptocurrencies Using Sentiment Analysis and Machine Learning." *Entropy* 21 (6): 589.

Vamossy, D. F., and R. Skog. 2023. "EmTract: Extracting Emotions from Social Media." Working paper, SSRN 3975884.

Vamvourellis, D., M. Toth, S. Bhagat, D. Desai, D. Mehta, and S. Pasquali. 2023. "Company Similarity Using Large Language Models." arXiv preprint arXiv:2308.08031.

van Zwam, M., A. Khalili, J. Jessurun, S. Oberoi, M. Beerepoot, S. Fernandez, J. Bijman, A. Easton, and I. Karatas. 2020. "Knowledge Graphs for Financial Services: The Path to Unlock New Insights from Your Data." White paper, Deloitte.

Vanetik, N., E. Podkaminer, and M. Litvak. 2023. "Summarizing Financial Reports with Positional Language Model." In *2023 IEEE International Conference on Big Data (BigData)*, pp. 2877–2883. IEEE.

Wan, Q., C. Wan, K. Xiao, R. Hu, D. Liu, and X. Liu. 2023. "CFERE: Multi-Type Chinese Financial Event Relation Extraction." *Information Sciences* 630: 119–134.

Wan, X., H. Deng, K. Zou, and S. Xu. 2024. "Enhancing the Efficiency and Accuracy of Underlying Asset Reviews in Structured Finance: The Application of Multi-Agent Framework." arXiv preprint arXiv:24 05.04294.

Wang, D., N. Raman, M. Sibue, Z. Ma, P. Babkin, S. Kaur, Y. Pei, A. Nourbakhsh, and X. Liu. 2023. "DocLLM: A Layout-Aware Generative Language Model for Multimodal Document Understanding." arXiv preprint arXiv:2401.00908.

Wang, S., Y. Pan, Z. Xu, B. Hu, and X. Wang. 2021. "Enriching BERT with Knowledge Graph Embedding for Industry Classification." In *Neural Information Processing*: *28th International Conference, ICONIP 2021*, Sanur, Bali, Indonesia, December 8–12, 2021, *Proceedings*, Part VI 28, pp. 709–717. Springer.

Wang, S., H. Yuan, L. M. Ni, and J. Guo. 2024. "QuantAgent: Seeking Holy Grail in Trading by Self-Improving Large Language Model." arXiv preprint arXiv:2402.03755.

Wang, S., H. Yuan, L. Zhou, L. M. Ni, H.-Y. Shum, and J. Guo. 2023. "Alpha-GPT: Human-AI interactive Alpha Mining for Quantitative Investment." arXiv preprint arXiv:2308.00016.

Wang, X., X. Pan, C. Chen, and J. Cui. 2023. "A Named Entity Recognition Model Based on BERT Model and Lexical Fusion in the Financial Regulation Field." In *2023 5th International Conference on Frontiers Technology of Information and Computer (ICFTIC)*, pp. 477–482. IEEE.

Wang, X., Y. Sun, C. Chen, and J. Cui. 2022. "A Relation Extraction Model Based on BERT Model in the Financial Regulation Field." In *2022 2nd International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI)*, pp. 496–501. IEEE.

Wang, Y., N. Lipka, R. A. Rossi, A. Siu, R. Zhang, and T. Derr. 2024. "Knowledge Graph Prompting for Multi-Document Question Answering." In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38, pp. 19206–19214.

Wang, Z., Y. Nie, P. Sun, N. H. Nguyen, J. Mulvey, and H. V. Poor. 2023. "ST-MLP: A Cascaded Spatio-Temporal Linear Framework with Channel-Independence Strategy for Traffic Forecasting." arXiv preprint arXiv:2308.07496.

Wang, Z., P. Sun, Y. Hu, and A. Boukerche. 2022. "A Novel Mixed Method of Machine Learning Based Models in Vehicular Traffic Flow Prediction." In *Proceedings of the 25th International ACM Conference on Modeling Analysis and Simulation of Wireless and Mobile Systems*, pp. 95–101.

Wen, Q., T. Zhou, C. Zhang, W. Chen, Z. Ma, J. Yan, and L. Sun. 2022. "Transformers in Time Series: A Survey." arXiv preprint arXiv:2202.07125.

Wimmer, C., and N. Rekabsaz. 2023. "Leveraging Vision-Language Models for Granular Market Change Prediction." arXiv preprint arXiv:2301.10166.

Wu, S., O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, and G. Mann. 2023. "BloombergGPT: A Large Language Model for Finance." arXiv preprint arXiv:2303.17564.

Xi, Z., W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou, et al. 2023. "The Rise and Potential of Large Language Model Based Agents: A Survey." arXiv preprint arXiv:2309.07864.

Xia, B., V. D. Rawte, M. J. Zaki, A. Gupta, et al. 2022. "FETILDA: An Effective Framework for Fin-Tuned Embeddings for Long Financial Text Documents." arXiv preprint arXiv:2206.06952.

Xie, Q., W. Han, Z. Chen, R. Xiang, X. Zhang, Y. He, M. Xiao, D. Li, Y. Dai, D. Feng, et al. 2024. "The FinBen: An Holistic Financial Benchmark for Large Language Models." arXiv preprint arXiv:2402.12659.

Xie, Q., W. Han, Y. Lai, M. Peng, and J. Huang. 2023. "The Wall Street Neophyte: A Zero-Shot Analysis of ChatGPT over Multimodal Stock Movement Prediction Challenges." arXiv preprint arXiv:2304.05351.

Xie, Q., W. Han, X. Zhang, Y. Lai, M. Peng, A. Lopez-Lira, and J. Huang. 2023. "PIXIU: A Large Language Model, Instruction Data and Evaluation Benchmark for Finance." arXiv preprint arXiv:2306.05443.

Xing, F. 2024. "Designing Heterogeneous LLM Agents for Financial Sentiment Analysis." arXiv preprint arXiv:2401.05799.

Xu, L., L. Zhu, Y. Wu, and H. Xue. 2024. "SuperCLUE-Fin: Graded Fine-Grained Analysis of Chinese LLMs on Diverse Financial Tasks and Applications." arXiv preprint arXiv:2404.19063.

Xu, Z., P. Zhou, X. Shi, J. Wu, Y. Jiang, B. Ke, and J. Yang. 2024. "FinTruthQA: A Benchmark Dataset for Evaluating the Quality of Financial Information Disclosure." arXiv preprint arXiv:2406.12009.

Xue, S., F. Zhou, Y. Xu, H. Zhao, S. Xie, C. Jiang, J. Zhang, J. Zhou, P. Xu, D. Xiu, et al. 2023. "WeaverBird: Empowering Financial Decision-Making with Large Language Model, Knowledge Base, and Search Engine." arXiv preprint arXiv:2308.05361.

Yadav, A., C. Jha, A. Sharan, and V. Vaish. 2020. "Sentiment Analysis of Financial News Using Unsupervised Approach." *Procedia Computer Science* 167: 589–598.

Yadav, D., S. Zhang, T. Jin, P. Krishnan, and D. Clarke. 2024. "Generative AI Based Virtual Assistant for Reconciliation Research." Working paper.

Yang, X., C. Zhang, Y. Sun, K. Pang, L. Jing, S. Wa, and C. Lv. 2023. "FinChain-BERT: A High-Accuracy Automatic Fraud Detection Model Based on NLP Methods for Financial Scenarios." *Information* 14 (9): 499.

Yekrangi, M., and N. Abdolvand. 2021. "Financial Markets Sentiment Analysis: Developing a Specialized Lexicon." *Journal of Intelligent Information Systems* 57: 127–146.

Yepes, A. J., Y. You, J. Milczek, S. Laverde, and L. Li. 2024. "Financial Report Chunking for Effective Retrieval Augmented Generation." arXiv preprint arXiv:2402.05131.

Yu, X., Z. Chen, Y. Ling, S. Dong, Z. Liu, and Y. Lu. 2023. "Temporal Data Meets LLM—Explainable Financial Time Series Forecasting." arXiv preprint arXiv:2306.11025.

Yu, Y., H. Li, Z. Chen, Y. Jiang, Y. Li, D. Zhang, R. Liu, J. W. Suchow, and K. Khashanah. 2024. "FinMem: A Performance-Enhanced LLM Trading Agent with Layered Memory and Character Design." In *Proceedings of the AAAI Symposium Series*, Vol. 3, pp. 595–597.

Yuan, H., S. Wang, and J. Guo. 2024. "Alpha-GPT 2.0: Human-in-the-Loop AI for Quantitative Investment." arXiv preprint arXiv:2402.09746.

Yuan, T., Z. He, L. Dong, Y. Wang, R. Zhao, T. Xia, L. Xu, B. Zhou, F. Li, Z. Zhang, et al. 2024. "R-Judge: Benchmarking Safety Risk Awareness for LLM Agents." arXiv preprint arXiv:2401.10019.

Yue, C., X. Xu, X. Ma, L. Du, H. Liu, Z. Ding, Y. Jiang, S. Han, and D. Zhang. 2023. "Leveraging LLMs for KPIs Retrieval from Hybrid Long-Document: A Comprehensive Framework and Dataset." arXiv preprint arXiv:2305.16344.

Yue, T., and D. Au. 2023. "GPTQuant's Conversational AI: Simplifying Investment Research for All." Working paper, SSRN 4380516.

Zeng, Z., W. Watson, N. Cho, S. Rahimi, S. Reynolds, T. Balch, and M. Veloso. 2023. "FlowMind: Automatic Workflow Generation with LLMs." In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pp. 73–81.

Zhang, B., H. Yang, and X.-Y. Liu. 2023. "Instruct-FinGPT: Financial Sentiment Analysis by Instruction Tuning of General-Purpose Large Language Models." arXiv preprint arXiv:2306.12659.

Zhang, B., H. Yang, T. Zhou, M. Ali Babar, and X.-Y. Liu. 2023. "Enhancing Financial Sentiment Analysis Via Retrieval Augmented Large Language Models." In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pp. 349–356.

Zhang, C., X. Liu, M. Jin, Z. Zhang, L. Li, Z. Wang, W. Hua, D. Shu, S. Zhu, X. Jin, et al. 2024. "When AI Meets Finance (Stockagent): Large Language Model-Based Stock Trading in Simulated Real-World Environments." arXiv preprint arXiv:2407.18957.

Zhang, K., O. Yoshie, and W. Huang. 2024. "BreakGPT: A Large Language Model with Multi-Stage Structure for Financial Breakout Detection." arXiv preprint arXiv:2402.07536.

Zhang, L., W. Cai, Z. Liu, Z. Yang, W. Dai, Y. Liao, Q. Qin, Y. Li, X. Liu, Z. Liu, et al. 2023. "FinEval: A Chinese Financial Domain Knowledge Evaluation Benchmark for Large Language Models." arXiv preprint arXiv:2308.09975.

Zhang, W., L. Zhao, H. Xia, S. Sun, J. Sun, M. Qin, X. Li, Y. Zhao, Y. Zhao, X. Cai, et al. 2024. "FinAgent: A Multimodal Foundation Agent for Financial Trading: Tool-Augmented, Diversified, and Generalist." arXiv preprint arXiv:2402.18485.

Zhang, X., R. Xiang, C. Yuan, D. Feng, W. Han, A. Lopez-Lira, X.-Y. Liu, S. Ananiadou, M. Peng, J. Huang, et al. 2024. "Dólares or Dollars? Unraveling the Bilingual Prowess of Financial LLMs Between Spanish and English." arXiv preprint arXiv:2402.07405.

Zhang, Y., K. Gong, K. Zhang, H. Li, Y. Qiao, W. Ouyang, and X. Yue. 2023. "Meta-Transformer: A Unified Framework for Multimodal Learning." arXiv preprint arXiv:2307.10802.

Zhang, Y., S. Mao, T. Ge, X. Wang, A. de Wynter, Y. Xia, W. Wu, T. Song, M. Lan, and F. Wei. 2024. "LLM as a Mastermind: A Survey of Strategic Reasoning with Large Language Models." arXiv preprint arXiv:2404.01230.

Zhang, Z., X. Bo, C. Ma, R. Li, X. Chen, Q. Dai, J. Zhu, Z. Dong, and J.-R. Wen. 2024. "A Survey on the Memory Mechanism of Large Language Model Based Agents." arXiv preprint arXiv:2404.13501.

Zhang, Z., Y. Sun, Z. Wang, Y. Nie, X. Ma, P. Sun, and R. Li. 2024. "Large Language Models for Mobility in Transportation Systems: A Survey on Forecasting Tasks." arXiv preprint arXiv:2405.02357.

Zhao, H., Z. Liu, Z. Wu, Y. Li, T. Yang, P. Shu, S. Xu, H. Dai, L. Zhao, G. Mai, et al. 2024. "Revolutionizing Finance with LLMs: An Overview of Applications and Insights." arXiv preprint arXiv:2401.11641.

Zhao, L., L. Li, X. Zheng, and J. Zhang. 2021. "A BERT Based Sentiment Analysis and Key Entity Detection Approach for Online Financial Texts." In *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pp. 1233–1238. IEEE.

Zhao, Q., J. Wang, Y. Zhang, Y. Jin, K. Zhu, H. Chen, and X. Xie. 2023. "CompeteAI: Understanding the Competition Behaviors in Large Language Model-Based Agents." arXiv preprint arXiv:2310.17512.

Zhao, W. X., K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al. 2023. "A Survey of Large Language Models." arXiv preprint arXiv:2303.18223.

Zhao, Y., Y. Long, H. Liu, L. Nan, L. Chen, R. Kamoi, Y. Liu, X. Tang, R. Zhang, and A. Cohan. 2023. "DocMath-Eval: Evaluating Numerical Reasoning Capabilities of LLMs in Understanding Long Documents with Tabular Data." arXiv preprint arXiv:2311.09805.

Zhao, Z. Y., Z. Zhu, G. Li, W. Wang, and B. Wang. 2023. "Generative Pretraining At Scale: Transformer-Based Encoding of Transactional Behavior for Fraud Detection." arXiv preprint arXiv:2312.14406.

Zhou, T., Z. Ma, Q. Wen, X. Wang, L. Sun, and R. Jin. 2022. “FEDformer: Frequency Enhanced Decomposed Transformer for Long-Term Series Forecasting.” In *International Conference on Machine Learning*, pp. 27268–27286. PMLR.

Zhou, T., P. Niu, L. Sun, R. Jin, et al. 2024. “One Fits All: Power General Time Series Analysis by Pretrained LM.” *Advances in Neural Information Processing Systems* 36: 43322–43355.

Zhou, W., S. Zhang, Y. Gu, M. Chen, and H. Poon. 2023. “UniversalNER: Targeted Distillation from Large Language Models for Open Named Entity Recognition.” arXiv preprint arXiv:2308.03279.

Zhu, D., J. Chen, X. Shen, X. Li, and M. Elhoseiny. 2023. “MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models.” arXiv preprint arXiv:2304.10592.

Zhu, J., S. Cai, F. Deng, and J. Wu. 2024. “Do LLMs Understand Visual Anomalies? Uncovering LLM Capabilities in Zero-Shot Anomaly Detection.” arXiv preprint arXiv:2404.09654.

Zhu, J., J. Li, Y. Wen, and L. Guo. 2024. “Benchmarking Large Language Models on CFLUE–A Chinese Financial Language Understanding Evaluation Dataset.” arXiv preprint arXiv:2405.10542.

Zmandar, N., A. Singh, M. El-Haj, and P. Rayson. 2021. “Joint Abstractive and Extractive Method for Long Financial Document Summarization.” In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pp. 99–105.

Zojaji, Z., R. E. Atani, A. H. Monadjemi, et al. 2016. “A Survey of Credit Card Fraud Detection Techniques: Data and Technique Oriented Perspective.” arXiv preprint arXiv:1611.06439.