Journal Of the Iraqia University (75-7) April (2026)

ISSN(Print): 1813-4521 OnLine ISSN:2663-7502
Journal Of the Iraqia University
available online at: https://www.mabdaa.edu.iq

large language models (LLMs) in Finance a state-of-the-art review of the last decade and Future directions

Mohamed Djafar Henni 1
1 Department of Economics, College of Business, Islamic University of Madinah, Saudi Arabia
mhenni@iu.edu.sa https://orcid.org/0009-0001-6099-9720

Abstract

The rapid rise of Generative Artificial Intelligence, including large language models (LLMs) such as GPT, BERT, and LLaMA, is reshaping financial markets by enabling large-scale analysis of unstructured textual information, which is central to pricing, risk management, and regulation. Despite growing empirical evidence on their predictive and operational value, existing studies remain fragmented, often separating methodological advances from financial and systemic implications. This study provides a systematic review of LLM applications in finance from 2015 to 2025, based on a structured search of Web of Science and Scopus databases. We synthesize evidence across key domains, including market prediction, sentiment analysis, portfolio management, auditing and fraud detection, regulatory technology, and ESG finance. Beyond applications, the review critically examines model reliability, limitations in numerical reasoning, bias, explainability, and governance challenges. Importantly, it highlights emerging systemic risks associated with coordinated LLM-driven financial decision-making. The study offers an integrated framework to support the responsible and resilient adoption of LLMs in financial systems.***Keywords:*** Large Language Models; Generative Artificial Intelligence;

Introduction

The rapid and profound emergence of **Generative Artificial Intelligence (GenAI)** and **Large Language Models (LLMs)**, including powerful transformer-based architectures like GPT, BERT, and Llama, is fundamentally transforming the financial sector by providing unprecedented capabilities in processing and interpreting vast, unstructured data, thereby reshaping the core functions of financial stability, risk management, and the coordination of market participants (Aldasoro et al., 2025; Eisfeldt & Schubert, 2025; Silva et al., 2024). The financial industry, which relies heavily on processing and aggregating information into price signals, has been particularly receptive to the advanced contextual understanding and high-fidelity content generation offered by LLMs (Haidar & Abbass, 2025). This technology is revolutionizing diverse areas, including corporate finance, asset pricing, household finance, and regulatory technology (Z. Feng et al., 2025; Mo & Ouyang, 2025).The transformative capabilities of LLMs are empirically validated across key financial applications, often surpassing traditional linguistic or machine learning approaches. In **Sentiment Analysis and Market Prediction**, specialized models like **FinBERT** and fine-tuned models such as GPT-3-based OPT or LLaMA-2 models extract nuanced sentiment from complex texts—ranging from financial news and corporate disclosures (like MD&A and 10-K filings) to social media discussions (such as WallStreetBets) and central bank speeches—to predict market movements, volatility, and returns (Alarnkar & Sankaranarayanan, 2025; Chiu & Hung, 2025; J. Fan et al., 2025; Huang et al., 2023; Iacovides et al., 2024; Ma et al., 2024; Nguyen et al., 2025). For instance, one strategy based on the GPT-3-based OPT model achieved an exceptional **Sharpe ratio of 3.05** and predicted stock market returns with an accuracy of 74.4% (Kirtac & Germano, 2024). LLMs are also critical in tasks such as **Information Extraction and Operational Efficiency**, where frameworks leveraging LLMs can achieve high accuracy rates, sometimes reaching **99.5 percent**, in extracting key financial indicators from unstructured PDF reports, thereby automating processes for accounting researchers, investors, and regulators (H. Li et al., 2023). Furthermore, LLMs enhance **Risk Management** by assisting in complex tasks like fraud detection, credit scoring through narrative data extraction, and quantifying hard-to-measure **novel risks** such, as ESG, geopolitical, and supply chain

disruption risks, from regulatory filings (S. Fan et al., 2025; Gadzinski & Vito, 2024; Gupta & Yan, 2025; H. Lu et al., 2025). In **Regulatory Interpretation**, models like GPT-4 are employed to distill complex documents, such as the Basel III capital requirements, into concise mathematical frameworks, thereby streamlining compliance (Z. Cao & Feinstein, 2024; Fazlija et al., 2025b). Moreover, LLMs demonstrate significant potential in **Investment Management**, not only by improving asset selection through economic insights but also by serving as sophisticated robo-advisors, with GPT-4 achieving a near-perfect 99% financial literacy score (J. H. Kim, 2023; Niszczota & Abbas, 2023; Schneider & Yilmaz, 2025).Despite this surge in demonstrable utility, the current body of research remains fragmented. Prior literature often exhibits a disjointed focus, either concentrating narrowly on the **financial or economic consequences** of AI without deep methodological scrutiny, or focusing predominantly on specialized **computer science methodologies**—such as novel model architectures, prompt engineering, fine-tuning strategies, or specialized data construction (e.g., multilingual financial datasets)—in isolation from their comprehensive financial impact (Dong et al., 2024; Kong et al., 2024). This dualistic focus prevents a holistic understanding of how cutting-edge LLM innovations translate into reliable, ethical practices in the high-stakes financial sector.This review directly **bridges this critical gap** by offering a rigorous, state-of-the-art synthesis that explicitly integrates the advancements in LLM methodology with their practical and strategic outcomes in finance, accounting, and investment management. This integrated approach is essential to address the pervasive cross-disciplinary challenges inherent in LLM adoption, including: 1) **Model Reliability and Bias**: LLMs, despite their eloquence, can suffer from **hallucination** and may exhibit intrinsic biases (such as the anchoring bias in forecasts) and ethical issues like racial bias in client communication, requiring sophisticated solutions like mechanistic interpretability and fairness frameworks. 2) **Technical Limitations and Reasoning**: LLMs often struggle with complex **numerical reasoning** and handling tabular data, motivating research into multi-agent reflection frameworks and specialized numeracy pre-training to enhance performance in quantitative tasks. 3) **Systemic Risk and Governance**: The potential for coordinated LLM-driven trading strategies introduces novel and understudied **systemic financial risk**, as evidenced by the immediate increase in systemic risk observed in Chinese banks following the launch of ChatGPT.By unifying the technological progression with its economic and risk implications, this review establishes a robust, multidisciplinary foundation for researchers and practitioners to navigate the opportunities and limitations of Generative AI, thereby promoting the responsible, trustworthy, and effective deployment of these models in shaping the future of the global financial ecosystem.

Search method procedure

This section provides a comprehensive overview of the methodology employed in sourcing the relevant academic papers necessary for conducting the current survey. We detail the specific search strategies, databases utilized, and inclusion criteria that guided our selection process.(TITLE-ABS-KEY (LLM* OR Large LANGUAGE model*) AND TITLE-ABS-KEY (bank*)) AND PUBYEAR > 2014 AND PUBYEAR < 2026 AND (LIMIT-TO (SUBJAREA , "ECON") OR LIMIT-TO (SUBJAREA , "BUSI")) AND (LIMIT-TO (LANGUAGE , "English"))

2.1. Search method

The primary platforms used to identify relevant literature for this review were the Web of Science and Scopus databases. A comprehensive search strategy was employed using an exhaustive list of keywords to capture the breadth of research at the intersection of Large Language Models (LLMs) and civil engineering. The search involved combinations of terms related to language models, including "Large Language Models", "LLMs", and specific prominent architectures such as "GPT", "BERT", "Transformer models", and "ChatGPT". These were paired with keywords covering various sub-disciplines and application areas within civil engineering, categorized broadly into:

- Building Information Modeling (BIM) and Design Automation
- Transportation and Traffic Management
- Geotechnical Engineering
- Risk Management and Safety
- Construction Management
- Structural Analysis, Design, and Optimization
- Building Codes and Regulations

Below are several examples of search queries that are presented in Table 1.

Table 1. Example Search Term Combinations Used in Web of Science.

LLM / NLP Term	Civil Engineering Term	Example Query Fragment
"Large Language Models"	"Construction Management"	TS=("Large Language Models" AND "Construction Management")
"LLMs"	"Geotechnical Engineering"	TS=("LLMs" AND "Geotechnical Engineering")
"GPT"	"Building Information Modeling" OR "BIM"	TS=("GPT" AND ("BIM" OR "Building Information Modeling"))
"BERT"	"Transportation Engineering"	TS=("BERT" AND "Transportation Engineering")
"Transformer models"	"Structural analysis" OR "Structural design"	TS=("Transformer models" AND ("Structural analysis" OR "Structural design"))
"chatGPT"	"Risk Analysis"	TS=("chatGPT" AND "Risk Analysis")

Note: Table 1 is truncated for brevity. The complete list of keyword combinations used in the search can be found in Appendix 1.

--- Deduplication Summary ---
Original WoS records found: 53
Original Scopus records found: 70
Total records before deduplication (sum of originals): 123
Total unique records after deduplication: 94
Success! Combined and deduplicated file saved as: combined_77_53.txt

More on the search criteria, the search timeframe was set from 2015 to 2025, encompassing the period of significant LLM development and early adoption. All document types were initially included. Subsequently, the retrieved results were filtered to exclude articles primarily focused on pure computer science advancements without clear civil engineering application and previously published review articles, ensuring the focus remained on primary research and novel applications within the target domain. The reference management software Mendeley was used to organize the selected literature. Figure 1 provides a summary of the search methodology utilized in this review.

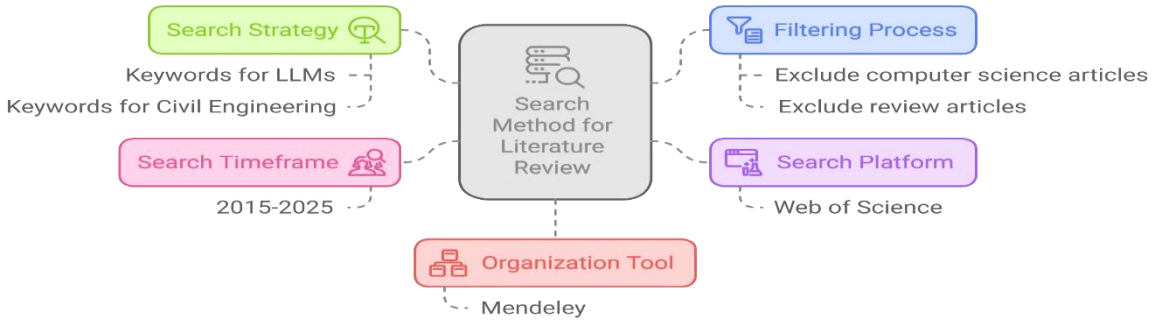

Figure 1. Search method to locate and retrieve academic articles and research papers.

2.2. Other reviews

A search through the existing literature revealed that while the application of Artificial Intelligence (AI) and Machine Learning (ML) in civil engineering sub-disciplines has been reviewed, comprehensive reviews focusing specifically on the broad application of modern Large Language Models (LLMs) across the entirety of civil engineering appear limited. Garcia et al (2022), for instance, reviewed the application of machine learning (ML) techniques within the construction industry, aiming to identify areas of application and future directions. The study analyzed studies published between 2015 and 2022 to assess the latest uses of ML in construction. The authors proposed a methodology that automatically identifies topics from article abstracts using the Bidirectional Encoder Representations from Transformers (BERT) technique, followed by manual selection of main topics. The review identified and analyzed relevant categories of ML applications in construction, including concrete technology, retaining wall design, pavement engineering, tunneling, and construction management. Additionally, the paper discussed various ML techniques, including supervised, deep, and evolutionary algorithms. The study aimed to provide future guidelines for researchers regarding ML applications in construction.Zhong & Goodfellow (2024) examined the application of the Transformer architecture and pre-trained Deep Learning models for Natural Language Processing (NLP) tasks within the Construction Management Systems (CMS) sector. The authors highlighted the increasing need for automated

methods in CMS and noted the under-explored potential of Transformer models in this domain. To address this, the study produced the first CMS domain corpora from academic papers and developed an end-to-end pipeline for pre-training and fine-tuning domain-specific Pre-trained Language Models. Four corpora were created, and transfer learning was used to pre-train BERT and RoBERTa models. The study found that the best-performing domain-specific models outperformed models pre-trained on general corpora when fine-tuned on two key NLP tasks: text classification for infrastructure condition prediction and named entity recognition for automatic construction control. Specifically, domain-specific pre-training improved F1 scores by 5.9% and 8.5% in these tasks, respectively. The authors concluded that these findings demonstrate the broader applicability of this approach to the Architecture, Engineering, and Construction sectors.Son et al (2025) presented a systematic literature review investigating how advanced technologies such as IoT, AI, digital twins, and optimization methods support smart transportation planning. The research examined the interrelationships between transportation challenges, proposed solutions, and enabling technologies to provide insights into smart mobility initiatives. Following PRISMA guidelines, the review identified 26 peer-reviewed articles published between 2013 and 2024 that examined smart transportation technologies. A Sentence BERT-based natural language processing approach was used to quantitatively assess relationships between key concepts by computing alignment scores among transportation challenges, technological solutions, and implementation strategies. The findings highlighted that real-time data collection, predictive analytics, and digital twin simulations significantly enhance traffic flow, safety, and operational efficiency while reducing environmental impacts. The analysis also revealed strong correlations between traffic congestion and public transit optimization, emphasizing the effectiveness of integrated, data-driven strategies. Furthermore, the study showed that IoT-based sensor networks and AI-driven decision-support systems play a crucial role in sustainable urban mobility by enabling proactive congestion management, multimodal transportation planning, and emission reduction strategies. From a policy perspective, the study emphasized the need for investment in urban-scale data infrastructures, the integration of digital twin modeling into long-term planning, and the alignment of optimization tools with public transit improvements. The study offered actionable recommendations for policymakers, engineers, and planners to guide data-driven resource allocation and legislative strategies for sustainable and technologically advanced transportation ecosystems.

1 Comparison between our survey and related surveys. Circles indicate areas covered but la e detail. Survey Financial LLMs Benchmarks Applications Challenges Lee et al. [2] Li et al. [4] Zhao et al. [5] ✓ ✓ ✗ ✗ ✓ ✗ ✗ ✓ ❍ ❍ ✓ ✓ ❍ ❍ ❍ ❍ Ours ✓ ✓ ✓ ✓

Large Language Models LLMs

The origins of language modeling lie in early computational linguistics and symbolic AI. In the 1950s, researchers such as Turing (1950) proposed theoretical frameworks for machines capable of mimicking human conversation. However, limited computational power and data availability constrained progress. Early rule-based systems, such as Joseph Weizenbaum's ELIZA (1966), relied on scripted responses rather than true language understanding. Later on, the 1980s saw the rise of statistical language models, which used probabilistic methods to predict word sequences. N-gram models, which estimate the likelihood of a word based on its preceding n-1 words, became a cornerstone of speech recognition and machine translation (Jelinek, 1976). These models, while effective for narrow tasks, lacked contextual depth and struggled with ambiguity.The resurgence of neural networks in the 2000s marked a paradigm shift. Bengio et al's seminal work (2003) on neural language models introduced architectures that could learn distributed representations of words. This laid the groundwork for word embeddings, dense vector representations capturing semantic relationships. In 2013, Tomas Mikolov and his colleagues (2013) at Google introduced Word2Vec, a highly efficient embedding technique that mapped words to vectors using shallow neural networks. Word2Vec's ability to analogize (e.g., "king – man + woman = queen") demonstrated the potential of neural methods for capturing linguistic patterns. Concurrently, GloVe (Global Vectors for Word Representation) by Stanford researchers provided a global context-aware embedding model (Pennington et al., 2014).The mid-2010s saw the integration of deep learning into NLP. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks enabled models to process sequential data with memory, thereby improving performance on tasks such as machine translation. The sequence-to-sequence (Seq2Seq) architecture, introduced by Google researchers (Sutskever et al., 2014), pairean encoder and a decoder to map input sequences to output sequences, revolutionizing tasks such as text summarization. However, RNNs faced limitations in capturing long-range dependencies due to vanishing gradients. The introduction of attention mechanisms addressed this by allowing models to focus on

relevant parts of the input dynamically (Bahdanau et al., 2014). This innovation culminated in the Transformer architecture, which replaced recurrence entirely with self-attention layers, enabling parallel processing and scalability (Vaswani et al., 2017).Several hybrid models bridged the gap between traditional and neural approaches. ELMo (Embeddings from Language Models), introduced by the Allen Institute for AI in 2018 (Peters et al., 2018), used bidirectional LSTMs to generate context-sensitive word embeddings. ELMo's dynamic representations outperformed static embeddings like Word2Vec, achieving state-of-the-art results on tasks such as question answering. That same year, the transformer revolution began, catalyzing the development of modern LLMs. Two landmark models emerged: Google's BERT (Bidirectional Encoder Representations from Transformers) and OpenAI's GPT (Generative Pre-trained Transformer). GPT-1 (Radford et al., n.d.) demonstrated the efficacy of unsupervised pre-training followed by task-specific fine-tuning. Trained on the BookCorpus dataset (7,000 unpublished books), GPT-1 used a unidirectional Transformer decoder to predict the next word in a sequence. Despite its 117 million parameters, GPT-1 achieved strong performance on tasks like text classification and entailment. BERT (Devlin et al., 2019), on the other hand, introduced bidirectional context by training on masked language modeling (MLM) and next sentence prediction (NSP). By masking 15% of input tokens and predicting them, BERT's 340 million parameters learned deep contextual relationships. BERT outperformed GPT-1 on benchmarks like GLUE (General Language Understanding Evaluation), achieving a score of 80.5% versus GPT-1's 72.8% (A. Wang et al., 2018).The success of early Transformers spurred a race to scale model size and data. In 2019, OpenAI released GPT-2, a 1.5-billion-parameter model trained on 40GB of web text. GPT-2's ability to generate coherent, contextually relevant text raised concerns about misuse, leading OpenAI to initially withhold the full model. Subsequently, GPT-3 marked a quantum leap with 175 billion parameters, trained on 570GB of text from Common Crawl, books, and Wikipedia (Brown et al., 2020). Its few-shot learning capabilities enabled users to prompt the model with minimal examples, achieving human-like performance in writing, coding, and reasoning. GPT-3's release sparked widespread adoption in applications like chatbots (e.g., ChatGPT's precursors) and content generation tools. By 2020, LLMs had been integrated into various domains, including Healthcare, where they assist in medical documentation and literature reviews (Benjamens et al., 2020), Education, where they automate feedback and provide personalized tutoring (Peng et al., 2020), and Engineering (Y. Wang et al., 2020).

Overview on the number of publications on LLMs in civil engineering

This section organizes and provides an overview of the publications identified concerning the application of Large Language Models (LLMs) in civil engineering, based on the search methodology described in Section 2. The search returned 53 publications from Web of Science and 70 from Scopus, yielding 123 records prior to deduplication. Analysis of publication dates for these initially identified papers reveals a clear and accelerating trend of interest in this topic within the civil engineering domain, particularly from 2021 onwards, as illustrated in Figure 2. This demonstrates exponential growth in research output in the last three years.

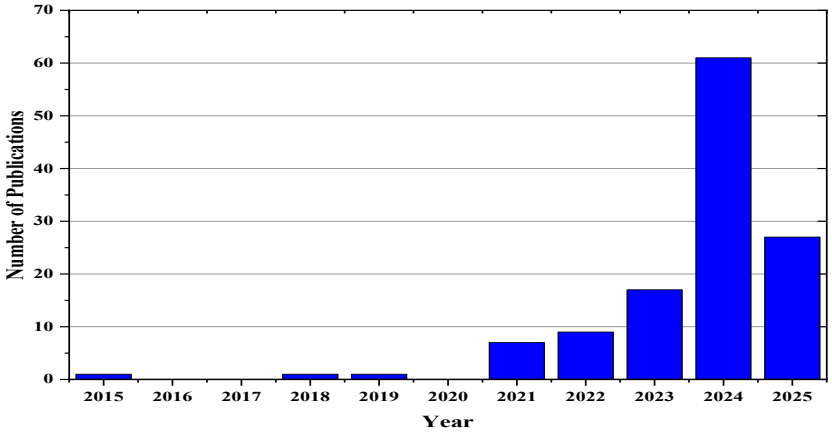

Figure 2. Total number of publications obtained by searching the keywords.

After applying the filtering criteria outlined in Section 2 (excluding pure computer science, reviews, etc.), a final set of 51 publications was selected for detailed review in this paper. These reviewed articles were categorized based on their primary civil engineering application area. The distribution across the main identified

categories is depicted in Figure 3. This breakdown indicates that, within the reviewed literature, applications related to BIM and Design Automation represent the most significant focus area to date, accounting for nearly half of the selected studies. Transportation, Geotechnical, Risk/Safety, and Construction Management/Education represent other key areas of application.

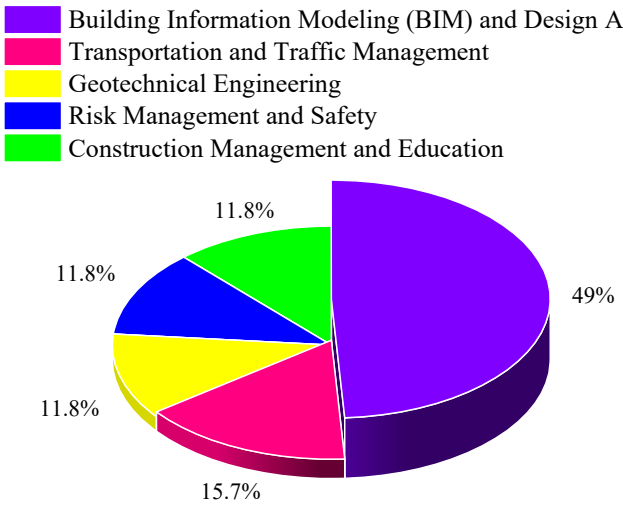

Figure 3. Percentage of publications corresponding to each subfield based on keyword searches.

An analysis of the publishers for the identified literature, as presented in Figure 4, reveals that Elsevier is the predominant outlet, with 41 papers, followed by MDPI with 21 papers and IEEE with 16 papers. Other notable publishers include Taylor & Francis (8), Wiley (7), Springer Nature (6), ASCE (9), and ACM (5).

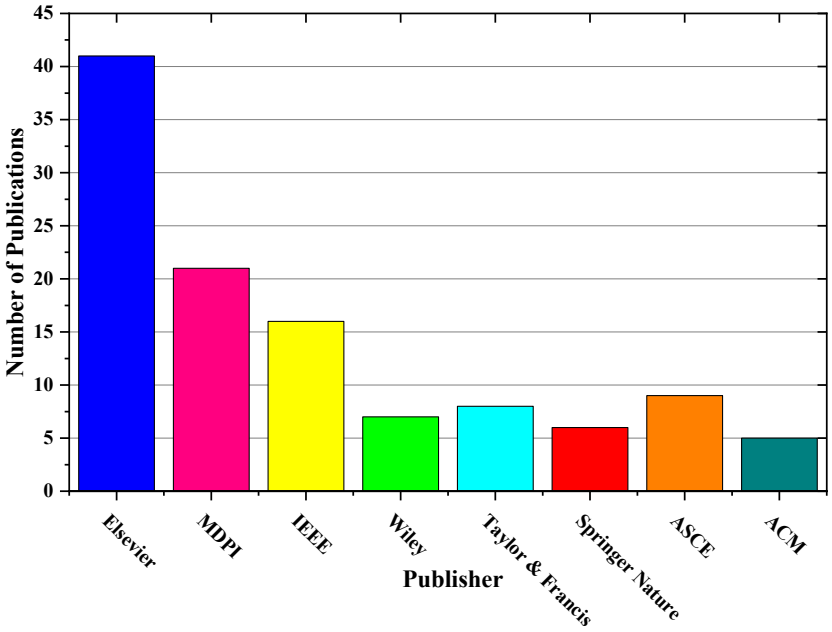

Figure 4. Number of publications per publisher.

Figure 5 illustrates a comprehensive network visualization of the co-occurrence analysis of keywords, highlighting how frequently different terms appear together in the literature. Each node on the network represents a specific keyword, while the connecting lines between the nodes indicate the strength of their co-occurrence, offering insights into thematic relationships within the subject area. In contrast, Figure 6 presents a detailed trend analysis of keyword usage over recent years, showcasing the fluctuations and emerging patterns in research focus.

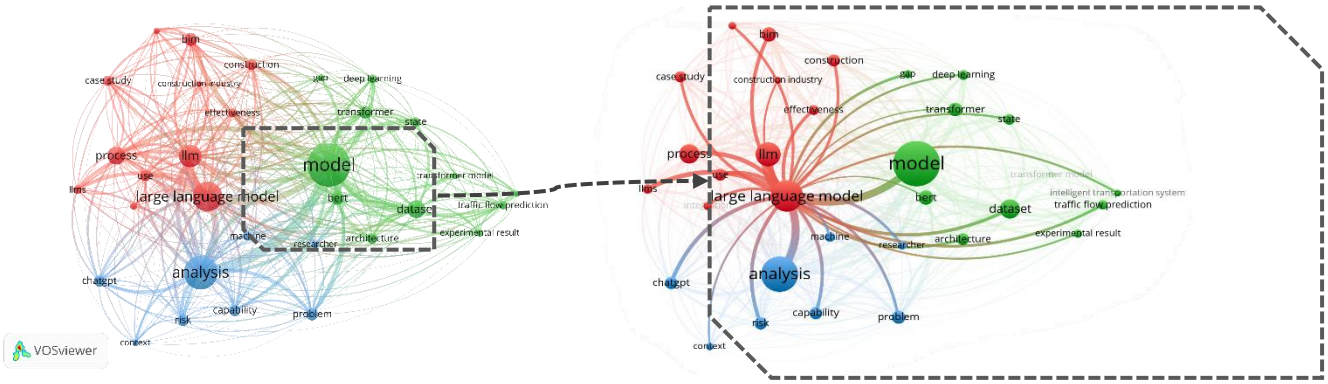

Figure 5. Co-occurrences of topic-related keywords with the key term “large language model”.

A keyword co-occurrence analysis, as illustrated in Figure 5, emphasizes the central themes within this research area. The terms "large language model" and "LLM" emerge as pivotal hubs, closely associated with terms such as "model," "analysis," "BERT," "transformer," and "ChatGPT." Additionally, there are strong connections to specific application domains, including building information modeling (BIM), the construction industry, transportation, architecture, and risk. This suggests a focused research concentration on core LLM technologies, their analytical capabilities, and their applications in particular contexts within civil engineering.Further analysis using keyword trend visualization (Figure 6) reveals insights into the recent evolution of research topics within this domain. While foundational concepts like "deep learning" appear as earlier terms (represented in purple tones) in this recent dataset, the latest trends, "Large Language Model" (shown in yellow tones), indicate a shift towards innovative implementations. Application domains like "BIM", "construction industry", and "traffic flow prediction" also emerge strongly in the latest trends, often explored through "case study" methodologies. This suggests the field is rapidly moving from establishing foundational models to exploring sophisticated integration and evaluating the effectiveness of these models in real-world civil engineering contexts.

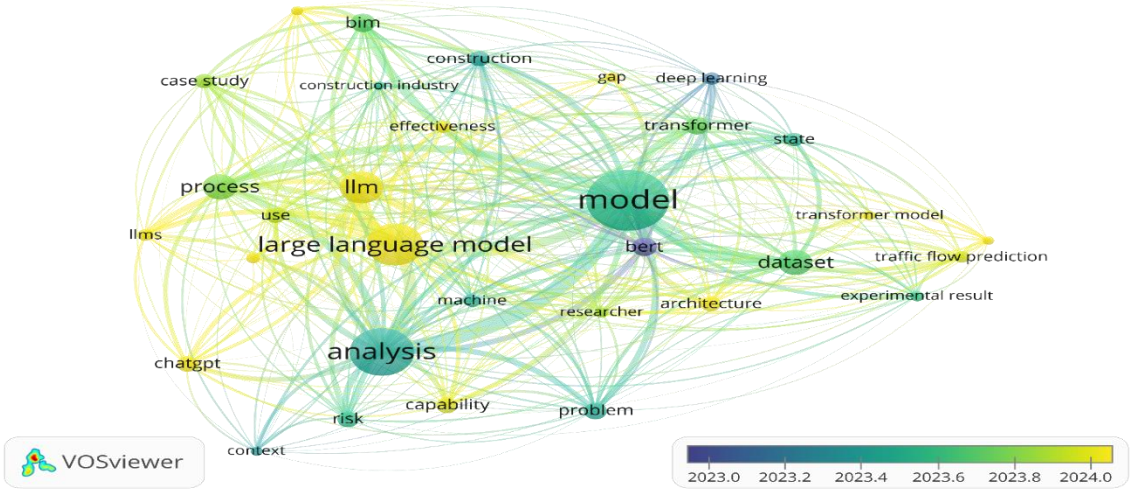

Figure 6. Analysis of trending keywords and their related co-occurrences.

LLM Architectures and Domain Adaptation

A. Core Architectures

The core architecture of this LLM includes FinBERT, a BERT-based large language model adapted for financial texts such as annual reports and 10-Ks, relying on a transformer architecture to effectively process complex financial language. Its core mechanism enables accurate analysis of news and market disclosures by capturing contextual nuances beyond traditional models (Araci, 2019; Y. Yang et al., 2020). FinBERT uses a stack of 12 transformer encoder layers, each consisting of a multi-head self-attention and feed-forward network, which has been pre-trained on 4.9 billion tokens from a financial corpus including 10-K, 10-Q, earnings transcripts, and analyst reports. This bidirectional setup processes the entire sequence in parallel, generating contextual embeddings for tokens in financial text (Araci, 2019; Y. Yang et al., 2020). Additionally, self-attention calculates query (Q), key (K), and value (V) vectors from the input embeddings through learned projections, then derives attention scores as scaled dot products of Q and K. This score weights the V vector, allowing each

token to pay attention to all other tokens regardless of distance, which is crucial for connecting distant phrases in 10-K reports or lengthy reports. Multi-head attention performs this in parallel across 12 heads, combining the outputs to capture various dependencies such as risk factors that refer to previous business descriptions (Ruan et al., 2021).On the other hand, in financial documents, self-attention excels at modeling long-range dependencies, such as connecting forward-looking statements in the MD&A section with risk disclosures separated by several pages, unlike RNNs which are constrained by sequential processing. Position encoding added to the embedding preserves sequence order, allowing the model to weigh relevance across thousands of tokens in annual reports or news streams. This mechanism improves tasks such as sentiment analysis in FiQA or PhraseBank datasets, where context spans sentences (Almarzoog et al., 2025; J. Kim et al., 2023; Liu et al., 2021; Mavillonio, 2024; Ruan et al., 2021). Therefore, conceptually, FinBERT's ability to build bidirectional contextual representations enables the identification of latent financial signals that are not easily captured by dictionary-based or traditional regression approaches, such as subtle shifts in risk tone or narrative inconsistencies in corporate disclosures.Despite its strengths in language modeling, FinBERT inherits BERT's limitations in numerical reasoning, as numbers are represented as discrete tokens without an understanding of magnitude or scale, which limits its use in financial ratio analysis or quantitative projections. These limitations have driven the development of numerically aware variants of BERT, such as BERT-M and BERT-V, which explicitly encode magnitude and numerical values into embeddings, enabling better integration between narrative text and quantitative information in financial documents.BERT-M focuses on scale pre-training, training the model to predict the scale of numbers relative to context, while BERT-V emphasizes value comparison, enabling better encoding of numerical dependencies with surrounding text. The combination of the two, BERT-M/V (or BERT-MV), combines these strengths, outperforming FinBERT by an average of 10.88% on the numeracy benchmark (F. Feng et al., 2021). Financial documents are dense with numbers representing revenue, ratios, and forecasts, where numeracy facilitates precision tasks such as risk assessment, sentiment analysis on earnings calls, or key metric extraction from reports. BERT-M/V improves accuracy in these areas with a better understanding of numerical semantics, supporting applications in Islamic finance analysis or sustainable investment evaluation where quantitative ethics are important (Choe et al., 2023; F. Feng et al., 2021). This improvement in numeracy supports downstream NLP tasks critical for governance and compliance in sectors such as energy or halal tourism financing.However, BERT-M/V remains encoder-only and non-generative, limited to classification or extraction without generating new text, sequences, or calculations like modern autoregressive LLM models. This model also lacks large-scale training like GPT-series models (billions of parameters on trillions of tokens), limiting generalization to complex reasoning or extensive financial corpora. These limitations make it inadequate for dynamic financial tasks requiring synthesis or scalability (Jin et al., 2021; Mahendra et al., 2025).Thus, FinBERT remains relevant as an efficient and stable encoder model for classification and signal extraction tasks, while generative models like FinLLaMA extend capabilities toward reasoning, summarization, and instruction-based decision-making, albeit with greater risks of hallucination and systemic implications. Albudairi et al (2024) and Konstantinidis et al (2024) argue that FinLLaMA, customized from Llama2 7B on financial data, enables instruction processing for summarization, reasoning, and trading signals, outperforming FinBERT by 44.7% in portfolio returns. These models handle complex tasks such as multi-step decision-making or news synthesis, which are vital for dynamic Islamic finance or governance analysis. Efficient parameter tuning such as LoRA keeps costs under control.

On the other hand, generative LLMs risk fabricating facts, leading to incorrect financial predictions or compliance issues in high-risk contexts. Unlike FinBERT's deterministic output, they reinforce biases from training data, posing systemic threats such as market disinformation. Mitigation through RAG adds overhead, reducing efficiency gains (Kang & Liu, 2023; Kirtac & Germano, 2025; Xu et al., 2026). Overall, the evolution from FinBERT to financial generative models reflects a shift from signal extraction to reasoning and synthesis, which enhances analytical capacity while increasing the need for model governance and systemic risk mitigation.

B. Domain Adaptation and Specialization

Fine-tuning, transfer learning, and model adaptation enable efficient customization of pre-trained financial language models such as FinBERT for specific languages such as Dutch (FinGEITje) or Japanese (JaFIn), utilizing an English-centric base to handle non-Latin scripts and local financial terminology. These techniques are essential for multilingual financial tasks such as sentiment analysis on regional reports or compliance checks in Islamic banking in diverse markets. Transfer learning initializes the model with general knowledge from a

pre-trained base (e.g., BERT), then adapts it through feature extraction or fine-tuning on target data, while maintaining computational efficiency (W. Lu et al., 2025). In this context, fine-tuning refers to adjusting parameters for a specific task, while model adaptation encompasses broader strategies such as cross-language learning and parameter-efficient module insertion. Model adaptation updates all or specific parameters on domain-specific datasets, such as Dutch financial reports for FinGEITje, improving accuracy by 5-15% on local sentiment tasks. Model adaptation for languages such as JaFIn uses fine-tuning on monolingual corpora, LoRA for low-dimensional updates, or adaptors to insert script-specific embeddings without full retraining (Ansell et al., 2023; W. Lu et al., 2025; Strangmann et al., 2024; Y. Yang et al., 2020). These methods address the global yet local nature of finance, enabling non-English models to accurately process earnings calls or ESG reports, while reducing training costs by 90% compared to models from scratch, ensuring scalability for low-resource languages while maintaining stability in classification tasks over generative risks. Implications include faster deployment in regulatory contexts (J. He et al., 2022; Kalluri, 2023; W. Lu et al., 2025).However, data scarcity in low-resource languages risks causing overfitting, which degrades zero-shot transfer (e.g., JaFIn struggles with rarely used kanji financial terms). Cultural and syntactic mismatches cause up to a 20% performance drop compared to monolingual training, reinforcing bias in ethical finance evaluation. Parameter-efficient methods such as adapters limit expressiveness for complex reasoning, while full adaptation requires high computation and, in particular, can hinder real-time Islamic governance applications (Han et al., 2021; J. He et al., 2022; Kalluri, 2023; Snegha et al., 2025; R. Zhang et al., 2020). Thus, language adaptation techniques expand the inclusivity of cross-jurisdictional financial LLMs, but demand a balance between efficiency, local accuracy, and systemic robustness.

C. Advanced Methodological Frameworks

The increasing complexity of financial applications makes it difficult for LLMs to handle them due to issues such as hallucinations, numerical inaccuracies, and lack of traceability, making them inadequate for the demands of accuracy and high-risk regulatory compliance. These limitations make financial systems vulnerable to risks that lead to audit and regulatory oversight failures (Chen et al., 2025; Kamble et al., 2025; Y. Li et al., 2024; Mateega et al., 2025). This has driven the development of advanced methodological frameworks that integrate LLMs with external components and control mechanisms. These approaches may include Retrieval-Augmented Generation (RAG), multi-agent systems, and code generation and reasoning, which together shift LLMs from language processing tools to more structured and auditable financial decision support systems. Retrieval-Augmented Generation (RAG) addresses this gap by integrating external knowledge retrieval with LLMs, basing output on verified financial data to reduce hallucinations and improve contextual accuracy in tasks such as compliance checking and time series forecasting (Kulpa & Wojarnik, 2025; Tailor, 2025; J. Wang et al., 2025). Multi-agent systems further enhance this by deploying collaborative LLM agents—such as analysts, critics, and risk managers—that discuss insights, verify outputs, and manage errors, improving performance in trading, portfolio optimization, and anomaly detection (Cruz, 2025; Kulpa & Wojarnik, 2025; Y. Yu et al., 2024). Code generation and reasoning mechanisms translate regulatory text into executable logic, enabling auditable compliance automation with structured workflows for verification and reporting (Cruz, 2025; Fazlija et al., 2025a; S. Li et al., 2026).This integration transforms LLMs from mere language processors into powerful, auditable financial decision support systems, providing traceable reasoning, error detection, and regulatory alignment through hybrid architectures. For example, multi-agent RAG settings achieve up to 23% higher precision in financial analysis while offering complete audit trails (J. Wang et al., 2025). Overall, these approaches enable scalable and compliant applications in volatile markets, surpassing single-model foundations in returns, risk management, and interpretation (Cruz, 2025; Fatemi & Hu, 2024; Kulpa & Wojarnik, 2025).

Core Applications in Financial Markets and Analysis

A. Prediction and Time Series Forecasting

Large language models (LLMs) play a crucial role in predicting and forecasting financial time series by complementing traditional numerical models through the extraction of previously unstructured textual information. In the context of stocks and indices, LLMs predict stock and index returns by processing unstructured textual data such as news, earnings reports, and social media to extract sentiment and semantic features correlated with price movements. These models, such as GPT-4 and its refined variants, generate embeddings or scores from news headlines and reports, which are then integrated into forecasting frameworks to outperform traditional methods in terms of accuracy and portfolio returns (T. Guo & Hauptmann, 2024; Lopez-lira et al., 2025; Okada et al., 2025; Siddique et al., 2025).

Furthermore, LLMs can also predict volatility and Value-at-Risk (VaR) by analyzing multi-source data including earnings calls and market news to model risk dynamics over a specific timeframe. Frameworks such as RiskLabs in the research by Cao et al (2025) use LLM for multimodal fusion of textual and time-series data, enabling effective prediction of market variance and volatility through feature extraction from qualitative sources. Risk-adjusted metrics developed through LLMs, such as AlphaSharpe, improve generalization in volatile conditions, demonstrating 3x higher predictive power for future risk-returns compared to benchmarks (Yuksel et al., 2025).Furthermore, for alternative assets such as FX, commodities, and crypto, LLM leverages sentiment from real-time news and technical indicators to generate trading signals and price forecasts. In cryptocurrency, models such as enhanced LLMs predict Ethereum prices with superior MSE and RMSE in few-shot settings, while multi-agent systems handle Bitcoin volatility through verbal feedback for 30% higher returns (Makri et al., 2025). Portfolio construction across crypto, commodities, and FX benefits from LLMs enhanced with RAG, improving Sharpe ratios and VaR estimates through comprehensive asset metrics (Hajaghaie & Thulasiram, 2025; Makri et al., 2025; Singhi, 2025).

B. Sentiment Analysis and Behavioral Finance

In sentiment and behavioral finance analysis, LLMs mark a shift from dictionary-based approaches to contextual modeling capable of capturing linguistic nuances such as sarcasm, negation, and inter-sentence framing. LLMs extract market sentiment from news, social media such as tweets, and analyst reports by processing large amounts of unstructured text through advanced natural language processing techniques. These models apply lexicon-based scoring, machine learning classifiers, and transformer architectures such as BERT to detect polarity (positive, negative, neutral) and intensity, often integrating multimodal data such as stock prices to improve accuracy (Agarwal et al., 2025; Heydarian et al., 2024; Varija & Hegde, 2024). Hybrid deep learning frameworks, including Convolutional Neural Networks (CNN) for feature extraction and Long Short-Term Memory (LSTM) for sequential dependencies, achieve up to 77% accuracy in linking sentiment signals with market trends across various sources (Dubey & Mahara, 2024).On the other hand, central bank communications, such as FOMC and ECB speeches, undergo tone analysis through LLMs that measure hawkish-dovish stance and forward guidance through semantic embedding and attention mechanisms. The refined model delineates policy nuances, economic outlook, and uncertainty levels, correlating tone shifts with yield curve reactions and volatility spikes. This approach reveals predictive power in monetary surprises, where negative tone often precedes risk-off movements in equity and currency markets (Y. Cao et al., 2025; J. Yang et al., 2025; Yuksel et al., 2025).Furthermore, LLMs enable large-scale studies of cognitive biases, overreactions, and emotional contagion among market participants, although exposure to noise and information manipulation remains a major challenge. Behavioral impacts manifest as anchoring biases in LLM-generated forecasts, where models over-rely on recent textual anchors such as headline extremes, distorting return predictions. Overreaction appears in LLM teams simulating crowd dynamics, amplifying short-term noise into exaggerated volatility estimates, while emotional contagion spreads through social media diffusion simulations, mimicking retail investor behavior in crypto and meme stocks (K. Guo & Xie, 2024; Moradi-Kamali et al., 2025; Zhao et al., 2023).

C. Investment and Portfolio Management

In the context of investment and portfolio management, large language models (LLMs) primarily serve as decision support tools by generating investor views, sentiment analysis, or predictive insights, rather than replacing human asset managers or established classical optimization models such as mean-variance analysis. These models excel at processing unstructured data such as news or earnings transcripts to inform views, but require integration with rigorous frameworks to ensure stability and risk-adjusted performance (Abe et al., 2024; Y. Lee et al., 2025; Mantshimuli & Mwamba, 2025; Saha et al., 2025).LLMs enhance portfolio processes through tasks such as return forecasting or style identification, but they still face limitations in standalone use due to market regime sensitivity and a lack of built-in optimization capabilities. For example, LLM-generated predictions improve when combined across personas but still perform poorly in buy-and-hold strategies in downturns without complementary strategies. This positions LLMs as supplementary tools for asset managers, aiding complex reasoning while deferring final allocation to experts (Abe et al., 2024; Hwang et al., 2025; Y. Lee et al., 2025; Saha et al., 2025).Furthermore, the Black-Litterman Model incorporates LLM output as investor views with confidence levels, addressing mean-variance sensitivity by blending market equilibrium with sentiment or forecasts derived from LLM (Y. Lee et al., 2025). Research by Mantshimuli & Mwamba (2025) shows that aggregating multi-LLM sentiment through LSTM, incorporated into Black-Litterman, yields superior Sharpe ratios compared to benchmarks on the S&P 500 portfolio. This integration reduces LLM biases

such as hallucinations, producing robust weights without replacing the model's core mathematics.Then, in factor-based allocation, LLMs combine quantitative factors (e.g., value, growth) with news representations to predict returns, often through careful fusion or mixed models for multimodal synergy (T. Guo & Hauptmann, 2025). This supports stock selection and risk management but relies on traditional optimization for portfolio construction, as LLMs alone would fail in complex regimes. Empirical testing shows that LLM-enhanced factor models outperform static methods in sector allocation, underscoring their supporting role in dynamic frameworks (Hossaina et al., 2025; Quek et al., 2025). On the other hand, LLM also enables the development of more personalized robo-advising by considering investors' risk profiles, ethical preferences, and long-term goals through natural narrative processing. However, limitations such as hallucination risk, lack of fiduciary accountability, and regulatory uncertainty restrict the direct use of LLM in automated investment decision-making.

Corporate and Regulatory Applications

A. Data Extraction and Management

In the context of data extraction and management, LLMs play an important role in automating the processing of complex and unstructured corporate documents, such as PDF-based annual reports, narrative disclosures, and sustainability reports. This addresses the challenge of processing visually complex PDFs, where traditional methods struggle with preserving layout and semantic relationships. Frameworks that combine LLMs with semantic hierarchy trees further improve efficiency for ad-hoc SQL queries on template-based document collections (Balsiger et al., 2024; Bansal et al., 2025; H. Li et al., 2023).Furthermore, integrating LLMs with structured reporting standards such as XBRL also facilitates validation, normalization, and cross-company and cross-jurisdiction comparisons. LLMs parse XBRL reports by labeling numbers with appropriate labels through instruction tuning, enabling extreme classification in financial filings regardless of domain-specific complexity. This generative approach outperforms rule-based parsers on long hybrid documents, extracting structured data from earnings transcripts and regulatory filings through generation enriched with retrieval. These capabilities support measurable analysis of SEC 10-K filings, assessing company performance based on metrics such as sustainability and innovation (Daimi & Iqbal, 2024; Khatuya et al., 2024; Sarmah et al., 2023; Yue et al., 2024). Furthermore, to measure non-financial disclosures, LLMs excel at extracting ESG data from company reports, including environmental and governance narratives that are often buried in unstructured text or tables. Techniques such as few-shot learning and fine-tuning enable models to generate analytical insights, such as risk assessments from sustainability sections, bridging textual and numerical data (Dimmelmeier et al., 2024; H. Li et al., 2023; Lin et al., 2024; X. Zhang & Wang, 2025). This automation reduces the need for manual curation, which is particularly important for research in finance, especially Islamic finance and sustainable investing, where non-financial metrics align with ethical frameworks

B. Auditing, Fraud Detection, and Lending

In auditing, fraud detection, and credit assessment, LLM is used to extract latent signals from management texts such as the MD&A section, financial statement footnotes, and auditor opinions. LLM improves fraud detection in auditing by analyzing latent signals in the Management Discussion and Analysis (MD&A) section through embeddings such as FinBERT and FinGPT, combined with LSTM for temporal fraud prediction. These models capture sentiment and uncertainty developing in textual disclosures, outperforming traditional word frequency methods on 30 years of financial reports. This approach integrates non-financial narratives to flag deviations, improving accuracy in financial statement audits (Bakumenko et al., 2024; Ergun & Sefer, 2025; Z. Zhang et al., 2022).In credit assessment, Xia et al (2025) explain that LLM extracts predictive features from narrative data to predict loan defaults, treating textual content as more than just meaningless chatter with added value beyond the sample. The model generates default probability scores from borrower narratives, improving predictions when added to structured features in FinTech datasets. This narrative augmentation refines risk profiles beyond numerical metrics, aiding ethical lending in the context of Islamic finance.On the other hand, LLM assesses audit opinions by processing unstructured reports for anomaly detection and compliance insights, supporting intelligent audit systems with dynamic risk evaluation (Yao et al., 2024). Tools such as AuditLLM enable multiprobe analysis of model outputs for governance audits, while embeddings detect latent fraud signals in ledgers (Amirizaniani et al., 2024). This capability simplifies the formation of opinions on financial health, aligning with regulatory requirements for transparency (Bakumenko et al., 2024; Mökander et al., 2023).C. ESG and Sustainability Finance

The application of LLM in ESG and sustainable finance marks a shift from rigid score-based assessments to dynamic narrative analysis by extracting contextual insights from sustainability reports, enabling real-time sentiment tracking and stakeholder-specific interpretation. Traditional static scores overlook nuanced disclosures, while LLM uses data-enriched generation to measure evolving risks from textual data, driving adaptive investment strategies. This evolution supports ethical frameworks such as those found in Islamic sustainable finance, prioritizing qualitative depth over quantitative rigidity (Bronzini et al., 2024; Ni et al., 2023; Wu et al., 2025; Zou et al., 2023).LLMs measure ESG risk and sentiment through refined models that process unstructured reports, generating highly precise structured scores on environmental and governance factors. Specialized tools such as ESGReveal and SusGen-GPT generate actionable metrics from disclosures, revealing temporal dynamics in reporting quality (Wu et al., 2025; Zou et al., 2023). For climate disclosures, LLMs assess TCFD compliance using models such as ClimateBERT refined on ClimaText, automating the extraction of transition and physical risks from financial reports (Domínguez-quiñones & Aliende, 2025). The framework detects greenwashing and scenario gaps, enhancing transparency in energy sector filings (Capetz et al., 2023). This narrative-based approach aligns with global standards, dynamically quantifying Scope 1-3 emissions narratives (Garrido-merch & Gonz, 2023; Lai & Chen, 2024). On the other hand, LLM also measures the impact of AI on green transformation by analyzing sustainability reports for innovation signals, such as AI-driven emission reductions in the supply chain. These tools evaluate how AI integration improves ESG performance, tracking contributions to net-zero goals through sentiment on technology-enabled decarbonization (Domínguez-quiñones & Aliende, 2025; Wu et al., 2025).

Challenges, Risks, and Governance Implications

A. Reliability and Technical Limitations

In terms of reliability and technical limitations, LLMs face a number of serious challenges ranging from bias, accuracy, and decision clarity, which undermine trust in high-risk decisions such as audits or sustainable finance assessments. These limitations stem from imbalanced training data and opaque architecture, requiring safeguards for ethical implementation, particularly in the context of Islamic finance where fairness aligns with maqasid al-shariah.Representation bias in LLMs arises from skewed training corpora, leading to unfair financial advice that disadvantages underrepresented groups or regions (Birti et al., 2025). Detection methods reveal disparities in credit rating outputs, where models reinforce historical prejudices in MD&A narrative analysis. Fairness audits, including demographic parity checks, uncover these issues, which require bias removal techniques such as adversarial training to ensure fair ESG risk assessments (Birti et al., 2025; Kong et al., 2024; Saha et al., 2025).On the other hand, hallucinations also haunt LLMs, generating fabricated financial metrics or ESG disclosures during numerical reasoning tasks, with failure rates of up to 60% on complex FinanceQA benchmarks (Mateega et al., 2025). Numerical difficulties arise in loan default prediction, where models fail at multi-step calculations from XBRL data despite having good narrative capabilities. Further tone manipulation erodes reliability, as models inadvertently alter sentiment in climate reports, misrepresenting TCFD compliance risks (Y. Guo & Yang, 2024; H. Lee et al., 2024).Furthermore, mechanistic interpretability dissects LLM attention mechanisms to track decision paths in fraud detection, revealing how embedding influences audit opinions. LLMs generate human-readable explanations through thought chain triggering, aiding transparency in sustainable investment strategies, although post-hoc methods like SHAP remain crucial for validating outputs. These XAI approaches bridge black-box limitations, supporting regulatory compliance in green financial transformation (Deng et al., 2025; Kong et al., 2024; Saha et al., 2025).

B. Ethical, Legal, and Societal Concerns

In addition to technical challenges, ethical, legal, and social issues are also becoming more prominent with the adoption of LLM in the financial sector. These issues challenge regulatory compliance and trust in sectors such as Islamic finance, where data governance is aligned with Sharia principles of privacy and fairness. In terms of privacy and data security, financial LLMs process sensitive customer data for tasks such as credit assessment and fraud detection, exposing vulnerabilities to data security breaches. Federated learning frameworks such as DPFedBank enable collaborative training without centralizing data, but communication costs and statistical heterogeneity remain threats (P. He et al., 2024; Yaramolu, 2025). Regulations demand encryption and zero-trust models, but AI integration amplifies the risk of unauthorized access in the banking sector.Furthermore, generative LLMs also risk violating intellectual property by reproducing copyrighted financial reports or companies' ESG datasets during training, complicating compliance in sustainable finance tools. The legal framework lags behind, with algorithmic trading and automated advisory systems facing scrutiny due to non-

transparent sources of market narratives (Sakit, 2024; Sekar, 2025). Additionally, malicious attacks leverage LLMs to generate synthetic phishing narratives or manipulate MD&A sentiment for fraud, further increasing systemic risk in high-frequency trading. While LLMs aid detection through anomaly patterns, malicious adjustments enable sophisticated fraud, necessitating robust models with XAI for accountability (Awosika et al., 2023).

C. Regulatory and Systemic Risk

The widespread application of LLM for risky stock predictions generates uniform signals across companies, where aligned models trigger simultaneous selling or buying, creating exogenous shocks such as sharp declines or market bubbles. Covariance analysis across LLMs from different regions reveals high prediction correlations in sectors such as technology, exacerbating contagion during stress events (Mcclellan, 2025). In addition, regulatory mechanisms must measure LLM covariance and enforce diversity in model architecture, data sources, and risk limits to prevent herd behavior. Data-driven governance includes stress testing for cultural and regional biases, along with real-time monitoring of alpha mining results (L. Cao, 2025; Mahdavi et al., 2025; Mcclellan, 2025). Furthermore, LLMs improve efficiency through rapid narrative synthesis of news and filings, enhancing alpha discovery in quantitative strategies such as Chain-of-Alpha (L. Cao, 2025). However, trend-following behavior reduces information diversity, inflating bubbles as seen in agent-based simulations of aversion-biased trading. The net effect benefits short-term liquidity gains but risks long-term stability, necessitating hybrid human-AI trading desks (Henning et al., 2025; Vidler & Walsh, 2025; H. Yu et al., 2025).

Conclusion and Future Research Directions

5. Conclusion

This review comprehensively explored the integration and applications of Large Language Models (LLMs) in civil engineering over the past decade, highlighting how these transformative technologies are reshaping research and practice across diverse subfields. Our survey revealed significant strides in leveraging LLMs for tasks ranging from automated design processes to advanced data analysis, while also identifying persistent challenges that require further investigation.In the domain of Building Information Modeling (BIM) and design automation, LLMs such as GPT and BERT have been pivotal in enabling natural language interfaces, automated compliance checking, and intelligent design detailing. Innovations such as virtual BIM assistants and ontology-driven question-answering systems have substantially reduced barriers to accessing and interpreting complex BIM datasets, paving the way for more efficient, user-friendly workflows. Generative AI models are further revolutionizing architectural and structural design by facilitating text-to-code translation and automating detailing tasks.In transportation and traffic management, LLMs have emerged as powerful tools for traffic flow prediction, addressing the spatio-temporal complexity inherent in urban mobility systems. Novel architectures like STGLLM-E and GPT4TFP have demonstrated superior forecasting accuracy by effectively modeling intricate correlations in traffic data. Additionally, explainable AI frameworks such as xTP-LLM are contributing to more transparent and interpretable traffic predictions, which are critical for informed urban planning and real-time traffic management.The application of LLMs in geotechnical engineering has primarily focused on educational tools, problem-solving, and numerical modeling support. Studies have illustrated how LLMs can assist in generating code for finite element analysis, recognizing nested named entities in technical texts, and facilitating knowledge transfer in academic contexts. While promising, these applications underscore the necessity of expert supervision to ensure the reliability and safety of engineering solutions derived from AI assistance.In risk management and safety, LLMs have proven effective at enhancing risk analysis through frameworks such as Human-In-The-Loop (HITL) systems, which integrate expert feedback to mitigate model limitations, such as hallucination and lack of domain specificity. Hybrid approaches combining open-source and closed-source LLMs have also demonstrated potential in secure data environments, particularly for housing defect management and safety assessments in high-risk industries such as mining and construction.In construction management and education, LLMs are contributing to innovations in automated grading, document generation, and advanced information retrieval. Systems like RAG4CM are transforming how professionals navigate vast volumes of project documents, while multimodal interfaces combining LLMs with virtual reality are revolutionizing human-robot collaboration in complex construction environments.In conclusion, the adoption of LLMs in civil engineering is rapidly progressing, with substantial achievements already evident across key subfields. However, realizing the full potential of these technologies hinges on responsible development, rigorous validation, and the integration of domain expertise to ensure accuracy, safety, and practical applicability. This review serves as both a synthesis of current advancements and a roadmap for future

research, guiding practitioners and researchers toward the strategic and ethical implementation of LLMs to innovate and optimize civil engineering practice.

6. Challenges and Future Directions

Despite the significant advancements, several challenges and future directions warrant consideration. Ensuring the accuracy and reliability of AI-generated output remains critical, especially in safety-sensitive applications. The need for expert oversight and validation of AI-driven processes is crucial to mitigate potential errors and ensure practical applicability. Further research is needed to refine LLM-BIM integration frameworks for handling more complex design scenarios and to develop more robust methods for semantic enrichment and knowledge integration across diverse domains within the AEC industry. Ethical considerations and best practices for the widespread adoption of AI in BIM workflows also require careful examination.

References

Abe, Y., Matsuo, S., Kondo, R., & Hisano, R. (2024). Leveraging Large Language Models for Institutional Portfolio Management: Persona-Based Ensembles. *2024 IEEE International Conference on Big Data (BigData)*, 4799–4808. https://doi.org/10.1109/BigData62323.2024.10825362

Agarwal, V., Rangaiah, Y. V., Biswal, S. K., N, R., Singh, P., & A, M. P. (2025). Applying Advanced Data Mining Technique to Predict Stock Market Trends from Textual Information. *2025 International Conference on Metaverse and Current Trends in Computing (ICMCTC)*, 1–5. https://doi.org/10.1109/ICMCTC62214.2025.11196740

Alarnkar, A. A., & Sankaranarayanan, K. G. (2025). Mind over market: Impact of investor sentiment on the Indian stock market. *Investment Management and Financial Innovations*, *22*(3), 273–292. https://doi.org/10.21511/imfi.22(3).2025.21

Albudairi, A. H. H., Kosov, M. E., & Staroverova, O. V. (2024). Analysis of the Main Changes in Financial Markets in the Context of Digitalization of the Economy. *International Journal of Financial Management and Economics*, *7*(2), 105–111. https://doi.org/https://doi.org/10.33545/26179210.2024.v7.i2.348

Aldasoro, I., Gambacorta, L., Korinek, A., Shreeti, V., & Stein, M. (2025). Intelligent financial system: How AI is transforming finance. *JOURNAL OF FINANCIAL STABILITY*, *81*. https://doi.org/10.1016/j.jfs.2025.101472

Almarzoog, K. H. A., Furaijl, H. B., Kadhum, N., Srayyih, F. H., Raheem, M. J., Faisa, F. G., Alshareef, H., Mohammed A, M., & Khalil, I. (2025). Advanced Financial Sentiment Analysis of Global Stock Market News Using Transformer-Based Deep Language Understanding Models. *2025 3rd International Conference on Cyber Resilience (ICCR)*, 1–8. https://doi.org/10.1109/ICCR67387.2025.11291723

Amirizaniani, M., Martin, E., Roosta, T., Chadha, A., & Shah, C. (2024). AuditLLM: A Tool for Auditing Large Language Models Using Multiprobe Approach. *CIKM '24: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (October 2024)*, 5174–5179. https://doi.org/10.1145/3627673.3679222

Ansell, A., Maria, E., Korhonen, A., & Vuli, I. (2023). Composable Sparse Fine-Tuning for Cross-Lingual Transfer. *ArXiv Preprint ArXiv*, 1–19.

Araci, D. T. (2019). FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. *ArXiv Preprint ArXiv*, *1908.10063*, 1–10.

Awosika, T., Shukla, R. M., & Pranggono, B. (2023). Transparency and Privacy: The Role of Explainable AI and Federated Learning in Financial Fraud Detection. *ArXiv Preprint ArXiv*, 1–9.

Bahdanau, D., Cho, K. H., & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. *International Conference on Learning Representations*. https://arxiv.org/abs/1409.0473v7

Bakumenko, A., Plant, C., & Hubig, N. C. (2024). Advancing Anomaly Detectio: Non-Semantic Financial Data Encoding with LLMs. *ArXiv Preprint ArXiv*, 1–9.

Balsiger, D., Dimmler, H., Egger-horstmann, S., & Hanne, T. (2024). Assessing Large Language Models Used for Extracting Table Information from Annual Financial Reports. *Computers*, *13*(257), 1–13. https://doi.org/https:// doi.org/10.3390/computers13100257

Bansal, R., Shah, B. M., Chanda, A., & Parate, V. (2025). Optimizing PDF Ingestion for Large Language Models in RAG Architectures. *International Journal OfAppliedMathematics*, *38*(3), 487–504. https://doi.org/https://doi.org/10.12732/ijam.v38i3s.163

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, *3*, 1137–1155.

Benjamens, S., Dhunnoo, P., & Meskó, B. (2020). The state of Artificial Intelligence-based FDA-Approved Medical Devices and Algorithms: An Online Database. *Npj Digital Medicine*, *3*(118), 1–8. https://doi.org/10.1038/s41746-020-00324-0

Birti, M., Maurino, A., & Osborne, F. (2025). Optimizing Large Language Models for ESG Activity Detection in Financial Texts. *ICAIF '25: Proceedings of the 6th ACM International Conference on AI in Finance*, 856–863. https://doi.org/10.1145/3768292.3770371

Bronzini, M., Nicolini, C., Lepri, B., Passerini, A., & Staiano, J. (2024). Glitter or Gold? Deriving Structured Insight from Sustainability Reports via Large Language Models. *ArXiv Preprint ArXiv*, 1–35.

Brown, T. B., Krueger, G., Mann, B., Askell, A., Herbert-voss, A., Winter, C., Ziegler, D. M., Radford, A., & Mccandlish, S. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Cao, L. (2025). Chain-of-Alpha: Unleashing the Power of Large Language Models for Alpha Mining in Quantitative Trading. *ArXiv Preprint ArXiv*.

Cao, Y., Chen, Z., Kumar, P., Pei, Q., Yu, Y., Li, H., Dimino, F., Ausiello, L., Subbalakshmi, K. P., & Ndiaye, P. M. (2025). RiskLabs: Predicting Financial Risk Using Large Language Model based on Multimodal and Multi-Sources Data. *Proceedings of International Workshop on Multimodal Financial Foundation Models (MFFMs) at 5th ACM International Conference on AI in Finance (MFFM Workshop @ ICAIF '24)*, 1–12.

Cao, Z., & Feinstein, Z. (2024). Large Language Model in Financial Regulatory Interpretation. *2024 IEEE SYMPOSIUM ON COMPUTATIONAL INTELLIGENCE FOR FINANCIAL ENGINEERING AND ECONOMICS, CIFER 2024*. https://doi.org/10.1109/CIFER62890.2024.10772991

Capetz, M., Chance, C., Pattichis, R., Vinella, A., Ghosh, R., & Chang, K.-W. (2023). Leveraging Language Models to Detect Greenwashing. *ArXiv Preprint ArXiv*, 1–7.

Chen, Z., Chen, J., Chen, J., & Sra, M. (2025). Standard Benchmarks Fail - Auditing LLM Agents in Finance Must Prioritize Risk. *ArXiv Preprint ArXiv*, 1–46.

Chiu, I.-C., & Hung, M.-W. (2025). Finance-specific large language models: Advancing sentiment analysis and return prediction with LLaMA 2. *PACIFIC-BASIN FINANCE JOURNAL*, *90*. https://doi.org/10.1016/j.pacfin.2024.102632

Choe, J., Noh, K., Kim, N., Ahn, S., & Jung, W. (2023). Exploring the Impact of Corpus Diversity on Financial Pretrained Language Models. *ArXiv Preprint ArXiv*, 2101–2112.

Cruz, A. D. La. (2025). Multi-Agent Large Language Models for Traditional Finance and Decentralized Finance. *Journal of Industrial Engineering and Applied Science*, *3*(1), 10–19. https://doi.org/https://doi.org/10.70393/6a69656173.323634

Daimi, S. A., & Iqbal, A. (2024). A Scalable Data-Driven Framework for Systematic Analysis of SEC 10-K Filings Using Large Language Models. *ArXiv Preprint ArXiv*, 1–10.

Deng, Y., Zhang, X., Zhou, D., Zhang, D., & Huang, B. (2025). Leveraging NLP in Finance: A Synergistic Approach Using Large Language Models and Chain-of-Thought Reasoning. *ICAICE '24: Proceedings of the 5th International Conference on Artificial Intelligence and Computer Engineering*, 494–500. https://doi.org/10.1145/3716895.3716983

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org/https://arxiv.org/abs/1810.04805v2

Dimmelmeier, A., Doll, H. C., Schierholz, M., Kormanyos, E., Fehr, M., Ma, B., Beck, J., Fraser, A., & Kreuter, F. (2024). Informing Climate Risk Analysis Using Textual Information – A Research Agenda. *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, 12–26. https://doi.org/https://doi.org/10.18653/v1/2024.climatenlp-1.2

Domínguez-quiñones, M., & Aliende, I. (2025). Assessment of TCFD Voluntary Disclosure Compliance in the Spanish Energy Sector: A Text Mining Approach to Climate Change Financial Disclosures. *World*, *6*(92), 1–30. https://doi.org/https://doi.org/ 10.3390/world6030092

Dong, M. M., Stratopoulos, T. C., & Wang, V. X. (2024). A scoping review of ChatGPT research in accounting and finance. *INTERNATIONAL JOURNAL OF ACCOUNTING INFORMATION SYSTEMS*, *55*. https://doi.org/10.1016/j.accinf.2024.100715

Dubey, A. S., & Mahara, P. S. (2024). Integrating Multimodal Deep Learning for Enhanced News Sentiment

Analysis and Market Movement Forecasting. *International Journal of Innovative Science and Research Technology*, *9*(6), 1290–1297. https://doi.org/https://doi.org/10.38124/ijisrt/IJISRT24JUN1691

Eisfeldt, A. L., & Schubert, G. (2025). Generative AI and Finance. *Annual Review of Financial Economics*, *17*(1), 363–393. https://doi.org/10.1146/annurev-financial-112923-020503

Ergun, Z. E., & Sefer, E. (2025). Financial Statement Fraud Detection via Large Language Models. *Intelligent Systems in Accounting, Finance, and Manajemen an International Journal*, *32*(4). https://doi.org/https://doi.org/10.1002/isaf.70021

Fan, J., Wei, X. Q., Yao, Y., Jiang, Y. B., Xin, C. K., & Feng, L. (2025). Computer-assisted rehabilitation system in the use of motor function recovery: A protocol for scoping review. *PLOS ONE*, *20*(7). https://doi.org/10.1371/journal.pone.0326865

Fan, S., Kong, D., Wu, Y., & Yu, H. (2025). Digital innovation and supply chain risk: A large language model-based analysis. *PACIFIC-BASIN FINANCE JOURNAL*, *92*. https://doi.org/10.1016/j.pacfin.2025.102799

Fatemi, S., & Hu, Y. (2024). Enhancing Financial Question Answering with a Multi-Agent Reflection Framework. *Proceedings of the 5th ACM International Conference on AI in Finance*, 1–8.

Fazlija, B., Ibraimi, M., Forouzandeh, A., & Fazlija, A. (2025a). Implementing Financial Regulations Using Large Language Models. *SSRN*, 1–30. https://doi.org/https://dx.doi.org/10.2139/ssrn.5010694

Fazlija, B., Ibraimi, M., Forouzandeh, A., & Fazlija, A. (2025b). Reasoning with financial regulatory texts via Large Language Models. *JOURNAL OF BEHAVIORAL AND EXPERIMENTAL FINANCE*, *47*. https://doi.org/10.1016/j.jbef.2025.101067

Feng, F., Rui, X., Wang, W., Cao, Y., & Chua, T.-S. (2021). Pre-training and Evaluation of Numeracy-oriented Language Model. *ICAIF '21: Proceedings of the Second ACM International Conference on AI in Finance*, *November 2021*, 1–9. https://doi.org/10.1145/3490354.3494412

Feng, Z., Hu, G., Li, B., & Wang, J. (2025). Unleashing the power of ChatGPT in finance research: opportunities and challenges. *FINANCIAL INNOVATION*, *11*(1). https://doi.org/10.1186/s40854-025-00770-3

Gadzinski, G., & Vito, L. (2024). ChatGPT: A canary in the coal mine or a parrot in the echo chamber? Detecting fraud with LLM: The case of FTX. *Finance Research Letters*, *70*. https://doi.org/10.1016/j.frl.2024.106349

Garcia, J., Villavicencio, G., Altimiras, F., Crawford, B., Soto, R., Minatogawa, V., Franco, M., Martínez-muñoz, D., & Yepes, V. (2022). Machine Learning Techniques Applied to Construction: A Hybrid Bibliometric Analysis of Advances and Future Directions. *Automation in Construction*, *142*(May), 104532. https://doi.org/10.1016/j.autcon.2022.104532

Garrido-merch, E. C., & Gonz, C. (2023). Fine-tuning ClimateBert Transformer with ClimaText for the Disclosure Analysis of Climate-Related Financial Risks. *ArXiv Preprint ArXiv*, 1–15.

Guo, K., & Xie, H. (2024). Deep Learning in Finance Assessing Twitter Sentiment Impact and Prediction on Stocks. *PeerJ Computer Science*, 1–19. https://doi.org/10.7717/peerj-cs.2018

Guo, T., & Hauptmann, E. (2024). Fine-Tuning Large Language Models for Stock Return Prediction Using Newsflow. *ArXiv Preprint ArXiv*, 1–16.

Guo, T., & Hauptmann, E. (2025). Exploring the Synergy of Quantitative Factors and Newsflow Representations from Large Language Models for Stock Return Prediction. *ArXiv Preprint ArXiv*, 1–30.

Guo, Y., & Yang, Y. (2024). EconNLI: Evaluating Large Language Models on Economics Reasoning. *ArXiv Preprint ArXiv*, 1–13.

Gupta, S., & Yan, H. (2025). Using Large Language Models to Estimate Novel Risk: Impact on Volatility. *JOURNAL OF PORTFOLIO MANAGEMENT*, *51*(7), 230–247.

Haidar, A., & Abbass, A. (2025). Navigating the Frontier of Finance: A Scoping Review of Generative AI Applications and Implications. *Generative Artificial Intelligence in Finance: Large Language Models, Interfaces, and Industry Use Cases to Transform Accounting and Finance Processes*, 215–252. https://doi.org/10.1002/9781394271078.ch12

Hajaghaie, A., & Thulasiram, R. K. (2025). Leveraging Large Language Models and Retrieval-Augmented Generation for Enhanced Multi-Asset Portfolio Construction. *2025 IEEE Symposium on Computational Intelligence for Financial Engineering and Economics (CiFer)*, 1–7. https://doi.org/10.1109/CiFer64978.2025.10975739

Han, W., Pang, B., & Wu, Y. (2021). Robust Transfer Learning with Pretrained Language Models through Adapters. *59th Annual Meeting Ofthe Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 854–861.

He, J., Berg-kirkpatrick, T., & Neubig, G. (2022). Towards a Unified View of Parameter-efficient Transfer Learning. *ArXiv Preprint ArXiv*, 1–15.

He, P., Lin, C., & Montoya, I. (2024). DPFedBank: Crafting a Privacy-Preserving Federated Learning Framework for Financial Institutions with Policy Pillars. *ArXiv Preprint ArXiv*, 1–11.

Henning, T., Ojha, S. M., Spoon, R., Han, J., & Camerer, C. F. (2025). LLM Agents do not Replicate Human Market Traders: Evidence from Experimental Finance. *ArXiv Preprint ArXiv*, 1–51.

Heydarian, P., Bifet, A., & Corbet, S. (2024). Understanding Market Sentiment Analysis: A Survey. *Journal of Economic Surveys*, 1125–1147. https://doi.org/10.1111/joes.12645

Hossaina, A., Ansaria, M. Q., Jeelanib, H., Digrac, M., & Syed, F. J. (2025). From Text to Returns: Using Large Language Models for Mutual Fund Portfolio Optimization and Risk-Adjusted Allocation. *ArXiv Preprint ArXiv*, 1–12.

Huang, A. H., Wang, H., & Yang, Y. (2023). FinBERT: A Large Language Model for Extracting Information from Financial Text*. *Contemporary Accounting Research*, *40*(2), 806–841. https://doi.org/10.1111/1911-3846.12832

Hwang, Y., Kong, Y., Zohren, S., & Lee, Y. (2025). Decision-informed Neural Networks with Large Language Model Integration for Portfolio Optimization. *ArXiv Preprint ArXiv*, 1–30.

Iacovides, G., Konstantinidis, T., Xu, M., & Mandic, D. (2024). FinLlama: LLM-Based Financial Sentiment Analysis for Algorithmic Trading. *5TH ACM INTERNATIONAL CONFERENCE ON AI IN FINANCE, ICAIF 2024*, 134–141. https://doi.org/10.1145/3677052.3698696

Jelinek, F. (1976). Continuous Speech Recognition by Statistical Methods. *Proceedings of the IEEE*, *64*(4), 532–556. https://doi.org/10.1109/PROC.1976.10159

Jin, Z., Jiang, X., Wang, X., Liu, Q., Wang, Y., Ren, X., & Qu, H. (2021). NumGPT: Improving Numeracy Ability of Generative Pre-trained Models. *ArXiv Preprint ArXiv*, 1–8.

Kalluri, K. (2023). Adapting LLMs for Low Resource Languages-Techniques and Ethical Considerations. *International Scientific Journal of Engineering and Management*, *2*(3), 1–11. https://doi.org/10.55041/ISJEM00140

Kamble, K., Russak, M., Mozolevskyi, D., Ali, M., Russak, M., & AlShikh, W. (2025). Expect the Unexpected: FailSafe Long Context QA for Finance. *ArXiv Preprint ArXiv*, 1–18.

Kang, H., & Liu, X.-Y. (2023). Deficiency of Large Language Models in Finance: An Empirical Examination of Hallucination. *ArXiv Preprint ArXiv*, 1–15.

Khatuya, S., Mukherjee, R., Ghosh, A., Hegde, M., Dasgupta, K., Ganguly, N., Ghosh, S., Goyal, P., Modeling, L., & Sachs, G. (2024). Parameter-Efficient Instruction Tuning of Large Language Models For Extreme Financial Numeral Labelling. *ArXiv Preprint ArXiv*, 1–13.

Kim, J. H. (2023). What if ChatGPT were a quant asset manager. *FINANCE RESEARCH LETTERS*, *58*(D). https://doi.org/10.1016/j.frl.2023.104580

Kim, J., Kim, H.-S., & Choi, S.-Y. (2023). Forecasting the S&P 500 Index Using Mathematical-Based Sentiment Analysis and Deep Learning Models: A FinBERT Transformer Model and LSTM. *Axioms*, *12*(9, 835), 1–22. https://doi.org/https://doi.org/10.3390/axioms12090835

Kirtac, K., & Germano, G. (2024). Sentiment trading with large language models. *FINANCE RESEARCH LETTERS*, *62*(B). https://doi.org/10.1016/j.frl.2024.105227

Kirtac, K., & Germano, G. (2025). Large Language Models in Finance: What is Financial Sentiment? *ArXiv Preprint ArXiv*, *March*, 1–15.

Kong, Y., Nie, Y., Dong, X., Mulvey, J. M., Poor, H. V., Wen, Q., & Zohren, S. (2024). Large Language Models for Financial and Investment Management: Models, Opportunities, and Challenges. *The Journal of Portfolio Management Quantitative Tools*, *51*(2), 211–231. https://doi.org/10.3905/jpm.2024.1.646

Konstantinidis, T., Iacovides, G., Xu, M., Constantinides, T. G., & Mandic, D. (2024). FinLlama: Financial Sentiment Classification for Algorithmic Trading Applications. *ArXiv Preprint ArXiv*, 1–6.

Kulpa, A., & Wojarnik, G. (2025). Prompt Engineering in Finance : An LLM-Based Multi-Agent Architecture for Decision Support. *European Research Studies Journal*, *XXVIII*(3), 1201–1217.

Lai, Y., & Chen, M.-Y. (2024). Explainable AI Approach to Construct a Human-Centric Consumer Application for Financial Climate Disclosures. *IEEE Transactions on Consumer Electronics*, *70*(1), 1112–1121. https://doi.org/10.1109/TCE.2023.3326953

Lee, H., Choi, Y., & Kwon, Y. (2024). Quantifying Qualitative Insights: Leveraging LLMs to Market Predict.

ArXiv Preprint ArXiv, 1–7.

Lee, Y., Kim, Y., Kim, J., Kim, S., & Lee, Y. (2025). LLM-Enhanced Black-Litterman Portfolio Optimization. *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10, 2025, Seoul, Republic of Korea*, *1*(1), 1–14. https://doi.org/10.48550/arXiv.2504.14345

Li, H., Gao, H. (Harry), Wu, C., & Vasarhelyi, M. A. (2023). Extracting Financial Data from Unstructured Sources: Leveraging Large Language Models. *JOURNAL OF INFORMATION SYSTEMS*, *39*(1), 135–156. https://doi.org/10.2308/ISYS-2023-047

Li, S., Chen, J., Yao, R., Hu, X., Zhou, P., & Qiu, W. (2026). Compliance-to-Code: Enhancing Financial Compliance Checking via Code Generation. *ArXiv Preprint ArXiv*, 1–16.

Li, Y., Wang, S., Ding, H., & Chen, H. (2024). Large Language Models in Finance : A Survey. *ArXiv Preprint ArXiv*, 1–11.

Lin, Y., Hulsebos, M., Ma, R., Shankar, S., Zeighami, S., Parameswaran, A. G., & Wu, E. (2024). Towards Accurate and Efficient Document Analytics with Large Language Models. *ArXiv Preprint ArXiv*, 1–16.

Liu, Z., Huang, D., Huang, K., Li, Z., & Zhao, J. (2021). FinBERT: A Pre-trained Financial Language Representation Model for Financial Text Mining. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 4513–4519.

Lopez-lira, A., Tang, Y., Chen, A., Davis, C., Eisfeldt, A., Han, X., Israelsen, R., Jiang, W., Lee, B., Jouanne-diedrich, H. Von, Naranjo, A., Ritter, J., Roussanov, N., Sheng, J., & Subra, A. (2025). Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models. *ArXiv Preprint ArXiv*, 1–142.

Lu, H., Zhang, Y., & Xu, J. (2025). Extraction of characteristic information from financial super-long texts and prediction of corporate violations. *RESEARCH IN INTERNATIONAL BUSINESS AND FINANCE*, *79*. https://doi.org/10.1016/j.ribaf.2025.103079

Lu, W., Luu, R. K., & Buehler, M. J. (2025). Fine-tuning Large Language Models for Domain Adaptation: Exploration of Training Strategies, Scaling, Model Merging and Synergistic Capabilities. *Npj Computational Materials*, *11*(84), 1–43. https://doi.org/10.1038/s41524-025-01564-y

Ma, F., Lyu, Z., & Li, H. (2024). Can ChatGPT predict Chinese equity premiums? *FINANCE RESEARCH LETTERS*, *65*. https://doi.org/10.1016/j.frl.2024.105631

Mahdavi, S., Chen, J. K., Joshi, P. K., & Huertas, L. (2025). Integrating Large Language Models in Financial Investments and Market Analysis: A Survey. *ArXiv Preprint ArXiv*, 1–21.

Mahendra, R., Spina, D., Cavedon, L., & Verspoor, K. (2025). Evaluating Numeracy of Language Models as a Natural Language Inference Task. *Findings Ofthe Association for Computational Linguistics: NAACL 2025*, 8351–8376.

Makri, E., Palaiokrassas, G., Bouraga, S., Polychroniadou, A., & Tassiulas, L. (2025). Ethereum Price Prediction Employing Large Language Models for Short-term and Few-shot Forecasting. *2025 7th Conference on Blockchain Research & Applications for Innovative Networks and Services (BRAINS)*, 1–10. https://doi.org/10.1109/BRAINS67003.2025.11302944

Mantshimuli, L. A., & Mwamba, J. W. M. (2025). Enhancing Portfolio Optimization with Multi-LLM Sentiment Aggregation: A Black-Litterman Integration Approach. *Investment Management and Financial Innovations*, *22*(3), 213–226. https://doi.org/10.21511/imfi.22(3).2025.16

Mateega, S., Georgescu, C., & Tang, D. (2025). FinanceQA: A Benchmark for Evaluating Financial Analysis Capabilities of Large Language Models. *ArXiv Preprint ArXiv*, 1–12.

Mavillonio, M. S. (2024). *Natural Language Processing Techniques for Long Financial Document* (No. 317). http://www.ec.unipi.it/ricerca/discussion-papers

Mcclellan, M. (2025). AI and Financial Fragility: A Framework for Measuring Systemic Risk in Deployment of Generative AI for Stock Price Predictions. *Journal OfRisk and Financial Managemen*, *18*(9), 1–40. https://doi.org/https://doi.org/10.3390/ jrfm18090475

Mikolov, T., Corrado, G., Chen, K., & Dean, J. (2013). Efficient Estimation ofWord Representations in Vector Space. *ArXiv Preprint ArXiv*, 1–12.

Mo, H., & Ouyang, S. (2025). (Generative) AI in Financial Economics. *JOURNAL OF CHINESE ECONOMIC AND BUSINESS STUDIES*, *23*(4), 509–587. https://doi.org/10.1080/14765284.2025.2569006

Mökander, J., Schuett, J., Kirk, H. R., & Jun, C. L. (2023). Auditing Largae Language Models: A Three-Layered Approach. *AI Ethics*, 1–36. https://doi.org/10.1007/s43681-023-00289-2

Moradi-Kamali, H., Rajabi-Ghozlou, M.-H., Ghazavi, M., Soltani, A., Sattarzadeh, A., & Entezari-Maleki, R.

(2025). Market-Derived Financial Sentiment Analysis: Context-Aware Language Models for Crypto Forecasting. *ArXiv Preprint ArXiv*, 1–13.

Nguyen, B. T., Chu, T. T., Ha, S. X., & Nguyen, A. T. (2025). Decoding market sentiment: the power of ChatGPT in explaining Bitcoin returns from X data. *CHINA FINANCE REVIEW INTERNATIONAL*. https://doi.org/10.1108/CFRI-05-2024-0278

Ni, J., Bingler, J., Colesanti-Senni, C., Kraus, M., Gostlow, G., Schimanski, T., Stammbach, D., Vaghefi, S. A., Wang, Q., Webersinke, N., Wekhof, T., Yu, T., & LeippolD, M. (2023). CHAT REPORT: Democratizing Sustainability Disclosure Analysis through LLM-based Tools. *ArXiv Preprint ArXiv*, 21–51.

Niszczota, P., & Abbas, S. (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*(A). https://doi.org/10.1016/j.frl.2023.104333

Okada, K., Nakasuji, M., & Tsukioka, Y. (2025). From Words to Returns: Sentiment Analysis of Japanese 10-K Reports Using Advanced Large Language Models. *PeerJ Computer Science*, *11*, 1–29. https://doi.org/10.7717/peerj-cs.3349

Peng, X., Han, C., Ouyang, F., & Liu, Z. (2020). Topic Tracking Model for Analyzing Student- Generated Posts in SPOC Discussion Forums. *International Journal of Educational Technology in Higher Education*, *17*(35), 1–22. https://doi.org/https://doi.org/10.1186/s41239-020-00211-4

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. *Proceedings Ofthe 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/https://doi.org/10.3115/V1/D14-1162

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2227–223. https://doi.org/https://doi.org/10.18653/v1/n18-1202

Quek, R., Heng, W., Vittori, E., Ong, K., Mao, R., Cambria, E., Mengaldo, G., Investments, A. I., Banking, I., Sanpaolo, I., & Science, D. (2025). Leveraging LLMs for Top-Down Sector Allocation in Automated Trading. *ArXiv Preprint ArXiv*, 1–15.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (n.d.). *Improving Language Understanding by Generative Pre-Training*. 1–12.

Ruan, S., Sun, X., Yao, R., & Li, W. (2021). Deep Learning Based on Hierarchical Self-Attention for Finance Distress Prediction Incorporating Text. *Computational Intelligence and Neuroscience*, *2021*(1165296), 1–11. https://doi.org/https://doi.org/10.1155/2021/1165296

Saha, P., Lyu, J., Saxena, A., Zhao, T., & Mehta, D. (2025). Large Language Model Agents for Investment Management: Foundations, Benchmarks, and Research Frontiers. *ICAIF '25: Proceedings of the 6th ACM International Conference on AI in Finance*, 736–744. https://doi.org/10.1145/3768292.3770387

Sakit, M. S. (2024). Regulating AI in Financial Services: Legal Frameworks and Compliance Challenges. *QANUN*, *08*(358), 30–39. https://doi.org/10.30546/2218-9130.026.2024.254

Sarmah, B., Zhu, T., Mehta, D., & Pasquali, S. (2023). Towards reducing hallucination in extracting information from financial reports using Large Language Models. *ArXiv Preprint ArXiv*, 1–5.

Schneider, C. J., & Yilmaz, Y. (2025). Stock portfolio selection based on risk appetite: Evidence from ChatGPT. *FINANCE RESEARCH LETTERS*, *82*. https://doi.org/10.1016/j.frl.2025.107517

Sekar, R. (2025). Designing Secure Data Applications and Products in the AI-driven Finance Sector. *World Journal of Advanced Engineering Technology and Sciences*, *15*(01), 556–563. https://doi.org/https://doi.org/10.30574/wjaets.2025.15.1.0238

Siddique, M. T., Jamee, S. S., Sajal, A., Mou, S. N., Mahin, R. H., Hossain, R., Hasan, M., & Chy, K. (2025). Enhancing Automated Trading with Sentiment Analysis: Leveraging Large Language Models for Stock Market Predictions. *The American Journal of Engineering and Technology*, *7*(03), 185–195. https://doi.org/10.37547/tajet/Volume07Issue03-16

Silva, L. C. E., Fonseca, G. de F., & Castro, P. A. L. (2024). Transformers and attention-based networks in quantitative trading: a comprehensive survey. *5TH ACM INTERNATIONAL CONFERENCE ON AI IN FINANCE, ICAIF 2024*, 822–830. https://doi.org/10.1145/3677052.3698684

Singhi, A. (2025). An Adaptive Multi-Agent Bitcoin Trading System. *ArXiv Preprint ArXiv*, *May*, 1–19.

Snegha, A., Sen, S., Pasi, P. S., Singhania, A., & Jyothi, P. (2025). Zero-Shot Cross-Lingual Transfer using

Prefix-Based Adaptation. *5th Workshop on Multilingual Representation Learning (MRL 2025)*, 385–396.

Son, H., Jang, J., Park, J., Balog, A., Ballantyne, P., Kwon, H. R., Singleton, A., & Hwang, J. (2025). Leveraging Advanced Technologies for (Smart) Transportation Planning: A Systematic Review. *Sustainability 2025, 17*(2245), 1–35. https://doi.org/https://doi.org/10.3390/su17052245

Strangmann, T., Purucker, L., Franke, J. K. H., Rapant, I., Ferreira, F., & Hutter, F. (2024). Transfer Learning for Finetuning Large Language Models. *ArXiv Preprint ArXiv*, 1–19.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. *Advances in Neural Information Processing Systems*, *4*(January), 3104–3112.

Tailor, P. (2025). Retrieval-Augmented Generation (RAG) for Real-Time Financial Market Analysis. *The American Journal of Interdisciplinary Innovations and Research*, *7*(07), 137–144. https://doi.org/10.37547/tajiir/Volume07Issue07-12

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind A Quarterly Review of Psychology and Philosophy*, *LIX*(236), 433–360. https://doi.org/https://doi.org/10.1093/MIND/LIX.236.433

Varija, B., & Hegde, N. P. (2024). An Automated Analytics Framework for Stock Trend Analysis from Multi-Modal Data. *SSRG International Journal of Electronics and Communication Engineering*, *11*(1), 116–130. https://doi.org/https://doi.org/10.14445/23488549/IJECE-V11I1P109

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*, 5999–6009. https://arxiv.org/abs/1706.03762v7

Vidler, A., & Walsh, T. (2025). Shifting Power: Leveraging LLMs to Simulate Human Aversion in ABMs of Bilateral Financial Exchanges, A bond market study. *ArXiv Preprint ArXiv*, 1–9.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 353–355. https://doi.org/https://doi.org/10.18653/v1/w18-5446

Wang, J., Ding, W., & Zhu, X. (2025). Financial Analysis : Intelligent Financial Data Analysis System Based on LLM-RAG. *ArXiv Preprint ArXiv*, 1–8. https://doi.org/https://doi.org/10.48550/arXiv.2504.06279

Wang, Y., Li, H., & Liao, L. (2020). DEM Construction Method for Slopes Using Three-Dimensional Point Cloud Data Based on Moving Least Square Theory. *Journal of Surveying Engineering*, *146*(3), 4020013. https://doi.org/10.1061/(ASCE)SU.1943-5428.0000320

Weizenbaum, J. (1966). ELIZA—A Computer Program for the Study of Natural Language Communication between Man and Machine. *Computational Linguistics*, *9*(1), 36–45. https://doi.org/10.1145/365153.365168

Wu, Q., Xiang, X., Huang, H., Wang, X., Jie, Y. W., Satapathy, R., Filho, R. S., & Veeravalli, B. (2025). SusGen-GPT: A Data-Centric LLM for Financial NLP and Sustainability Report Generation. *Findings Ofthe Association for Computational Linguistics: NAACL 2025*, 1184–1203.

Xia, Y., Shi, Z., Du, X., & Zheng, Q. (2025). Extracting Narrative Data via Large Language Models for Loan Default Prediction: When Talk isn't Cheap. *Applied Economics Letters*, *32*(4), 481–486. https://doi.org/10.1080/13504851.2023.2275647

Xu, X., Wen, F., Chu, B., Fu, Z., Lin, Q., & Liu, J. (2026). Finance-Specific Deployment of Large Language Models FinBERT2: A Specialized Bidirectional Encoder for Bridging the Gap in Finance-Specific Deployment of Large Language Models. *KDD '25: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, *2*(January), 5117–5128. https://doi.org/10.1145/3711896.3737219

Yang, J., Tang, Y., Li, Y., Zhang, L., & Zhang, H. (2025). Cross-Asset Risk Management: Integrating LLMs for Real-Time Monitoring of Equity, Fixed Income, and Currency Markets. *ArXiv Preprint ArXiv*, 1–9.

Yang, Y., UY, M. C. S., & Huang, A. (2020). FinBERT: A Pretrained Language Model for Financial Communications. *ArXiv Preprint ArXiv*, *2006.08097*, 1–5.

Yao, X., Wu, X., Li, X., & Xu, H. (2024). Smart Audit System Empowered by LLM. *ArXiv Preprint ArXiv*, 1–7.

Yaramolu, L. S. K. G. (2025). Privacy-Driven Federated AI in Financial Fraud Detection and Risk Scoring. *World Journal of Advanced Engineering Technology and Sciences*, *15*(1), 2492–2504. https://doi.org/https://doi.org/10.30574/wjaets.2025.15.1.0493

Yu, H., Li, F., & You, J. (2025). LiveTradeBench: Seeking Real-World Alpha with Large. *ArXiv Preprint ArXiv*, 1–37.

Yu, Y., Yao, Z., Li, H., Deng, Z., Jiang, Y., Cao, Y., Subbalakshmi, K., Xiong, G., He, Y., Huang, J., Li, D., & Xie, Q. (2024). FINCON: A Synthesized LLM Multi-Agent System with Conceptual Verbal Reinforcement for Enhanced Financial Decision Making. *ArXiv Preprint ArXiv*, 1–30. https://doi.org/https://doi.org/10.48550/arXiv.2407.06567

Yue, C., Xu, X., Du, L., Liu, H., Ding, Z., Jiang, Y., Han, S., & Zhang, D. (2024). Enabling and Analyzing How to Efficiently Extract Information from Hybrid Long Documents with LLMs. *ArXiv Preprint ArXiv*, 1–17.

Yuksel, K. A., Sawaf, H., & Jose, S. (2025). AlphaSharpe: LLM-Driven Discovery of Robust Risk-Adjusted Metrics. *ArXiv Preprint ArXiv*, 1–8.

Zhang, R., Reddy, R. G., Sultan, A., Castelli, V., Ferritto, A., Florian, R., Kayi, E. S., Roukos, S., Sil, A., & Ward, T. (2020). Multi-Stage Pre-training for Low-Resource Domain Adaptation. *2020 Conference on Empirical Methods in Natural Language Processing*, 5461–5468.

Zhang, X., & Wang, M. (2025). Large Language Models for Financial Knowledge Extraction Analytical Insights and Corporate Planning Support. *Mathematical Modeling and Algorithm Application*, *6*(2), 44–56. https://doi.org/https://doi.org/10.54097/7am6vk38

Zhang, Z., Ma, Y., & Hua, Y. (2022). Financial Fraud Identification Based on Stacking Ensemble Learning Algorithm: Introducing MD & A Text Information. *Computational Intelligence and Neuroscience*, *2022*(1), 1–14. https://doi.org/10.1155/2022/1780834

Zhao, Y., Du, Z., Xu, S., Cheng, Y., Mu, J., & Ning, M. (2023). Social Media, Market Sentiment and Meme Stocks. *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, 1197–1202. https://doi.org/10.1109/COMPSAC57700.2023.00181

Zhong, Y., & Goodfellow, S. D. (2024). Domain-Specific Language Models Pre-Trained on Construction Management Systems Corpora. *Automation in Construction*, *160*(September 2023), 105316. https://doi.org/10.1016/j.autcon.2024.105316

Zou, Y., Shi, M., Chen, Z., Deng, Z., Lei, Z., Zeng, Z., Yang, S., Tong, H., Xiao, L., & Zhou, W. (2023). ESGReveal: An LLM-based Approach for Extracting Structured Data from ESG Reports. *ArXiv Preprint ArXiv*, 1–17.

□