Generative-AI in Finance : Opportunities and Challenges of

Akshar Prabhu Desai *
Google
Mountain View, California, USA
akshard@google.com

Tejasvi Ravi *
Google
Mountain View, California, USA
ravitejasvi@google.com

Mohammad Luqman
Google
Mountain View, California, USA
moluqman@google.com

Ganesh Mallya
Google
Mountain View, California, USA
ganeshmallya@google.com

Nithya Kota
Google
Mountain View, California, USA
nithyakota@google.com

Pranjul Yadav
Google
Mountain View, California, USA
pranjulyadav@google.com

* These authors contributed equally.

Abstract—Gen-AI techniques are able to improve understanding of context and nuances in language modeling, translation between languages, handle large volumes of data, provide fast, low-latency responses and can be fine-tuned for various tasks and domains.

In this manuscript, we present a comprehensive overview of the applications of Gen-AI techniques in the finance domain. In particular, we present the opportunities and challenges associated with the usage of Gen-AI techniques. We also illustrate the various methodologies which can be used to train Gen-AI techniques and present the various application areas of Gen-AI technologies in the finance ecosystem.

To the best of our knowledge, this work represents the most comprehensive summarization of Gen-AI techniques within the financial domain. The analysis is designed for a deep overview of areas marked for substantial advancement while simultaneously pin-point those warranting future prioritization. We also hope that this work would serve as a conduit between finance and other domains, thus fostering the cross-pollination of innovative concepts and practices.

Index Terms—Large Language Models, Machine Learning, Payments, Finance, Gen-AI

I. INTRODUCTION

Financial sector has historically applied traditional machine learning technologies to address domain specific challenges. For example, clustering [1], predictive analysis [2], anomaly detection [3], time-series monitoring [4], and graph-based learning have found extensive use for finance-specific problems such as detecting trends, prediction, risk modeling, and better data representation. However, Generative AI (Gen-AI) technologies, such as large language models (LLMs), have a superior ability to understand tasks that involve language compared to any traditional techniques. Kashyap et al. [5] discuss that LLMs can effectively perform traditional machine learning tasks and, when used correctly, can even outperform traditional methods such as classification.

Gen-AI techniques are able to improve understanding of context and nuances in language modeling [6], translation between languages, handle large volumes of data, provide fast and low-latency responses and can be fine-tuned [7] for various tasks and domains. The advancement was sparked by the development of transformer architecture, which is based solely on attention mechanisms, dispensing with recurrence and convolutions [8]. Experiments have been performed to demonstrate how Gen-AI techniques have superior quality while requiring significantly less time to train [9, 10].

There are several opportunities associated with the usage of Gen-AI technologies in finance [11]. Broadly speaking, they can be used across multiple avenues such as interactive (e.g. chatbots), assistive (e.g. easing payment related activities, right card to use), educative (e.g. enabling users to understand finance concepts) and advisory (e.g. trading assistant).

Along with the opportunities, there lies several challenges associated with the widespread adoption. Some of these challenges originate from the scarcity of private and sensitive quality data available for training, issues stemming from pre-training and fine-tuning, inference latency, computational costs associated with deploying these models in production, exorbitant cost associated with the pricing of APIs available for commercial usage and the inherent biases embedded within these models [12, 13].

There are several avenues via which Gen-AI technologies can be trained and developed for financial use-case. Users can leverage either open source models or external service providers for their tasks [14]. Alternatively, one can explore the usage via prompt engineering i.e. zero-shot and few-shot learning. In scenarios, when utilizing Gen-AI techniques out of the box is not working for the task at hand, then one could leverage fine-tuning (e.g. instruction specific, task-specific or parameter efficient) to train in order to perform better at the specific tasks the user has in mind.

Gen-AI techniques also been widely applied in the finance industry with its application to customer service and support (e.g. sentiment analysis, chat-bots), text summarization and assistance (e.g. summaries from large text, recommending knowledge from large corpora), auto-filling forms, risk man-

arXiv:2410.15653v2 [cs.AI] 30 Oct 2024

agement (e.g. market risk analysis, credit risk monitoring, anomaly detection), investment trading (e.g. quantitative analysis) and document processing (e.g. regulatory compliance) [15] [16].

The organization of this paper is as follows. In Section II, we discuss the various opportunities associated with the usage of Gen-AI in the finance ecosystem. In Section III, we present the various challenges associated with the usage of these techniques. Further, in Section IV, we illustrate the various methodologies which can be used to train Gen-AI for widespread adoption. Furthermore, in Section V, we present the various application areas of Gen-AI in finance. Lastly, in Section VI, we conclude this manuscript.

II. Opportunities of Gen-AI in finance

In this section, we outline the opportunities associated with the usage of Gen-AI techniques in the finance domain. Broadly, we classify the opportunities into four major categories i.e. interactive, assistive, educative and advisory.

A. Interactive

Gen-AI technology has the potential to become a state-of-the-art technology for both Task-Oriented Dialog [17] and Open-Domain Dialog, thereby making it the go-to technology for building conversational and interactive applications. The associated inherent ability with these techniques to ask follow up questions, track references to previous entities, and user preferences makes them more coherent and consistent. They are a vast improvement over rule-based chat-bots which could respond to a fixed set of queries with a pre-configured set of responses. Fine-tuning and Retrieval Augmented Generation (RAG) [18] can further help Gen-AI techniques, become more domain specific and/or access more up to date relevant information to support customers ranging from routine to complex inquiries while creating more personalized and engaging experiences. McKinsey [19] estimates that Gen-AI could further reduce the volume of human-serviced contacts by up to 50 percent, depending on a company's existing level of automation. Further, Klarna, a Buy Now Pay Later (BNPL) company has observed that using Gen-AI techniques within its customer service tool can lead to a $40M impact [20].

Gen-AI techniques could also be used to create appealing backgrounds for debit, credit cards, coupons and gift cards. Many models including Parti [21] can generate high-fidelity photo-realistic images and supports content-rich synthesis involving complex compositions and world knowledge. Further, Gen-AI techniques can be used to create dynamic visually pleasing themes for apps that are more engaging and context specific.

B. Assistive

Gen-AI techniques have the ability to assist through digital automation, co-piloting, auto-filling and shopping.

In particular, agents using Gen-AI techniques, can be trained to execute APIs [22] thereby enabling them to execute tasks, automate routines thereby substantially improving the quality of life with humans having mobility and dexterity challenges. Further, a payments co-pilot can help identify opportunities for better payment options among cards, instrument modes (e.g. cards, bank accounts).

Gen-AI techniques are able to parse and understand page content and hence can help towards auto-filling, which is a complex problem due to the dynamic and changing nature of forms across the internet. Further, agents using Gen-AI technology can be used as digital assistants. LLaSA [23], a digital assistant is able to parse the product inventory and provide recommendations that closely match the user's needs.

C. Educative

Gen-AI techniques have the potential to create applications which can guide, educate and empower users w.r.t. financial literacy.

Gen-AI technology can be used to perform deep dives into existing financial data about a customer or company, thereby uncovering financial trends that could lead to cost efficiency measures, or investing opportunities. In particular, Gen-AI techniques have the ability to ingest vast amounts of financial data to analyze market sentiment, news and other data sources to help inform/create better trading strategies [24].

Further, chatbots [25] built using Gen-AI techniques can understand complex topics like mortgage planning and investments strategies (e.g. stock options, mutual funds).

D. Advisory

Financial institutions can leverage the capabilities of Gen-AI techniques to build powerful tools for financial advisory applications (e.g. risk assessment, pro-active detection, summarization and recommendation).

Gen-AI techniques can be used to build sophisticated machine learning models by processing vast amounts of financial data to identify anomalous patterns of fraud. Such models can be used for better credit scoring, risk assessment for loans, to identify fraudulent activities in payment systems and to forecast financial risks using chain of thought prompting [26] to elicit better reasoning and accuracy. Further, Gen-AI techniques have demonstrated a remarkable ability towards document summarization (i.e. concise yet informative) and identifying key themes.

III. Challenges of using Gen-AI in finance

Strict regulations and constraints within the financial domain hinders the development of Gen-AI applications. Further, Gen-AI's existing limitations present challenges for effective use in financial applications. In this section, we explore these challenges in more detail.

A. Data

1) Data Constraints: There is a dearth of publicly available financial datasets owing to the sensitive nature of financial data. This limits the availability of open financial Gen-AI models [27]. BloombergGPT, a proprietary Gen-AI model, is the only major model trained for financial tasks from

ground up. BloombergGPT is trained on 363 billion tokens of Bloomberg's own proprietary data and 345 billion tokens of publicly available data [28].

While many financial institutions have access to proprietary data, they still face data availability challenges due to the institutional inertia and privacy related concerns. Kruse et al. [29] conducted a survey of financial service experts to identify key challenges of using Gen-AI and two thirds of the experts indicated the lack of quality training data as the primary obstacle. Further, the same study [29] identified financial institutions' unwillingness to move the data to cloud as one of the major limiting factors in using Gen-AI in finance. Financial institutions see their data as their core-business and have concerns in moving this data to distributed data centers (or cloud) which is necessary for training.

The lack of quality training data and privacy concerns can be addressed using synthetic data [30]. Synthetic data is artificially generated data that mimics the statistical properties of the real world data. Since the data is artificially generated, it also helps alleviate the concerns around data privacy. Samuel et al. [30] provides a detailed overview of the current state of art of synthetic data generation in finance. However, they also highlight the challenges with synthetic data such as over-fitting and lack of open benchmarks.

2) Nature of Financial Data: Ljung et al. [31] showed that financial data is different from data in other domains because it does not follow Gaussian distribution and hence can not be normalized using state-of-the-art approaches for downstream modeling tasks. This requires the creation of new pre-processing methods for machine learning use-cases. Further, Samuel et al. [30] found that multi-modality and heterogeneity of financial data along with its intricate inter-dependencies are harder to be captured and modeled by Gen-AI technologies.

B. Challenges in fine tuning

Gen-AI models require substantial fine tuning efforts when they are applied to specific domains such as finance [32]. Fine-tuning is a technique to adapt a pre-trained model for a more specific purpose (e.g. financial applications). Li et al. [33] provided a comprehensive comparison of large language models (e.g. LLaMa) for financial applications and observed that fine tuned models can outperform generic models w.r.t financial tasks.

Further, FinBERT [34], which uses LoRA (Low Rank Adaptation) technique [35] to fine-tune open source LLMs, highlights the importance of pre-training along with fine tuning. In particular, they emphasized that financial vocabulary needs to be carefully embedded into the pre-training stages along with fine-tuning (e.g., associating "bank" with the word "lending" as compared to the word "river").

Li et al. [33] also highlights the critical nature of high quality human labeled data for fine tuning. They mentioned that, in the finance domain, cost of producing human labeled data is harder and expensive due to challenges around reg-ulations, privacy and availability of humans with specialized knowledge.

C. Computation costs

The two major contributors to computational costs are training cost and inference cost. Computational costs associated with training Gen-AI models from the ground up can easily run into billions of dollars [36]. However, Xia et al. [36] noted that fine tuning III-B can help reduce this cost for certain applications.

Inference cost associated with the integration of Gen-AI workflows will be high considering the billions of stock market trades, credit card payment and banking transactions within a day [37]. Further, pricing models of Gen-AI APIs available for commercial use vary from \$5 to \$15 per million input tokens. For example, if we assume that each credit card transaction involves 500 tokens, this translates to a cost of \$0.0025 per transaction or \$5 million for 2 billion transactions per day in additional expenses. Bryce et al. [38] provided an overview about how these pricing costs would impact Gen-AI deployment in finance.

D. Secondary Considerations

Additional considerations of using Gen-AI techniques in finance domain can arise from domain independent issues, embedded domain bias, privacy and regulatory concerns.

Domain independent issues of Gen-AI techniques such as hallucination and inconsistent reasoning [39] using incorrect information might lead to significant financial loss [40].

Embedded domain bias stems from the fact that different regions have different laws governing the kind of information, which can be used for making financial decisions (e.g. credit worthiness) [41]. As a result, Gen-AI techniques trained on such data might further perpetuate biases without being fully aware of their existence.

Data leaks due to sophisticated prompt engineering techniques [42] raises substantial privacy and regulatory concerns [12, 43] surrounding users private data. This complexity further increases due to new privacy regulations (e.g. right to be forgotten) adopted in many regions of Europe [44]. This increase in complexity can be attributed to the fact that a model might still preserve some aspects of the data as part of its learning, even though the underlying data is requested to be deleted.

IV. GEN-AI METHODOLOGIES

This section presents various ways to train or fine-tune Gen-AI techniques for potential use-cases in finance.

A. Out-of-Box

In this technique, the users leverage either open source models or Gen-AI service providers like OpenAI, Gemini, Microsoft and Perplexity for their tasks. This method requires no training data, and either minimal to no compute resources.

Further, users can make use of prompt engineering [45] techniques to complete their task. Broadly, there are two main

varieties of prompt engineering techniques i.e. zero-shot [46] and few-shot [47]. In the case of zero-shot learning, Gen-AI techniques are able to perform the task at hand without seeing any prior examples, but just by leveraging the knowledge present in them. On the other hand in case of few-shot learning, users provide a few examples as part of the prompt for the Gen-AI techniques to learn and perform the task at the end of the prompt. In particular, in this technique, the Gen-AI techniques can leverage their induction heads [48] to perform in-context learning [49] and arrive at the result.

B. Fine-tuning

In scenarios, where utilizing LLMs out of the box is not working for the task at hand, then users could leverage fine-tuning to train the LLMs to perform better at the specific tasks the user has in mind. At a high level, fine-tuning techniques can be divided into the following categories.

1) Instruction Fine-tuning: In instruction fine-tuning [50], the pre-trained LLMs are further trained on a labeled set of prompts and answer pairs. This newly trained LLM is tuned to answer specifically to specific kinds of instruction/prompt. As the name suggests, the training data need to be in the form of instructions. So users need to either collect the training data in the form of instructions or could leverage prompt template libraries like prompt-engine-py or dynamic prompts to take normal datasets and convert them into instruction datasets for fine-tuning.

2) Task specific Fine-tuning: This technique [51] involves the users fine-tuning a pre-trained Gen-AI model to perform a specific kind of task in mind. For example, the users might fine-tune the pre-trained LLM to perform sentiment detection given an input prompt. It involves very few examples for the LLM to train on, but still requires a decent amount of compute as the entire model needs to be loaded into memory for the training part.

Task specific fine-tuning [51] is prone to exhibit a phenomenon called catastrophic forgetting. In catastrophic forgetting, the underlying LLM has forgotten the knowledge of the world it had obtained as part of its pre-training and its performance on the other tasks after task specific fine-tuning is much worse than its performance on the same tasks before fine-tuning. In order to circumvent the issue of catastrophic forgetting, users can employ multi-task instruction fine-tuning or employ parameter efficient fine-tuning described in section IV-B3.

3) Parameter Efficient Fine-tuning: Instruction Fine-Tuning IV-B1 or Task specific fine-tuning IV-B2 are resource intensive and are plagued with catastrophic forgetting. In order to circumvent both these issues users can leverage Parameter Efficient Fine-tuning (PEFT) [52]. Two most commonly used techniques in PEFT are LoRA [35] which rely on re-parametrization technique and soft prompts which add additional trainable layers to the LLM.

1) LoRA introduces a new way to train the LLMs by retaining their pre-training knowledge. The weights of

the LLM are frozen and new low rank decomposition matrices are added to every layer in the transformer.
2) Soft Prompts [53] fall under the additive method paradigm where no weights of the model are changed. Instead of modifying the weights of the pre-trained LLMs, soft prompts rely on prepending trainable tokens or soft prompts to the input tokens during training for a given task. The loss is propagated all the way to these trainable tokens and an efficient representation is learnt for the task at hand. The idea of these soft tokens is to choose the right vectors for a given space on the N-dimensional hypersphere of the embedding vector space. Since these soft tokens are very light-weight, multiple such soft tokens can be learnt similar to LoRA.

C. Agentic Systems

Even though LLMs have made great strides in their ability to understand natural languages (via zero-shot or few shot examples), LLMs still struggle in providing accurate up to date information, hallucinate answers [54, 55] or are unable to perform precise mathematical calculations [56]. A very simple solution is to provide the ability for LLMs to utilize external tools such as search engine, calculators etc which help the LLMs in overcoming their major drawbacks.

D. Quantization

Traditionally models were trained and deployed using 32-bit floating point numbers to represent the parameters of the model. With the exponential increase in the number of parameters in LLMs, one major concern that arises is the amount of time taken for inference. In order to speed up the inference time researchers leverage quantization to represent the parameters of the model using less bits (float16, bfloat16 or int8) without much loss in accuracy. Users utilize one of the two quantization schemes

1) Post Training Quantization: In this method a pre-trained 32-bit floating point number based model is converted into low bit numbers. The quantization maybe data free or a very small amount of data called calibration data can be used. A crucial step in this quantization scheme is to find a good quantization ranges for the quantizer as noted in [57]. In this quantization scheme no retraining of the LLM is involved. This method of quantization (specifically the case with data free quantization) is particularly helpful when security or data privacy may limit data access [58].
2) Quantization aware training (QAT) involves in retraining the pre-trained LLM using training data. In this quantization scheme since the quantization operation is non-differentiable, during back-propagation the gradient is approximated using an identity function also known as Straight Through Estimator [59]. This method is particularly useful if the quantized model will be in use for a quite some time. Since this method involves retraining of the LLM on lower precision, the training must be

performed for an extended period so that quantized LLM converges to better loss in the loss manifold [58].

V. Gen-AI Applications in finance ecosystem

Gen-AI techniques are still in its nascent stages and its implementation within finance sector comes with its own set of risks and challenges. However, the transformative nature of this technology is driving many applications in this domain. In this section, we will highlight such applications both in the industry and the research community.

A. Numerical Reasoning

Several datasets consisting of complex numerical reasoning tasks have been proposed to evaluate the efficacy of Gen-AI techniques. For example, FinQA [60], is a dataset with Question-Answer pairs over Financial reports written by financial experts. FinQA, consist of questions such as "Considering the weighted average fair value of options, what was the change of shares vested from 2005 to 2006?". Similarly, ConvFinQA [61], an extension of FinQA, is a multi-turn conversational question-answering dataset over financial reports, consisting of 3,892 conversations with 14,115 questions. GPT-4 achieved an accuracy of 78% on FinQA and 76% on ConvFinQA dataset which is much higher than the average human [62]. This ability has enabled real world used cases like Fintool [63] - an AI equity research tool that is engineered to discover financial insights about companies.

Credit Karma [64], is making use of the Intuit Assist to launch Gen-AI experiences which include asking questions about a user's personal finance, understanding spend and personal financial roadmap.

B. Trading

Gen-AI techniques have also seen applications in trading yielding impressive results in research settings. In particular, Wu et al. (2023) introduced BloombergGPT [28], a 50-billion parameter Gen-AI technique, designed specifically for the financial domain. They demonstrated the effectiveness of their proposed technique for sentiment analysis, financial question answering, while maintaining proficiency in general language tasks.

Mai (2024) proposed StockGPT [65], which uses token sequences to learn predictive patterns via the attention mechanism there by demonstrating substantial impact. Further, Lopez et al. [66] showcased that incorporating advanced Gen-AI techniques into the investment decision-making process can yield accurate predictions and enhance the performance of quantitative trading strategies. Furthermore, Fatourosa et al. [67] proposed that a portfolio rebalanced monthly using the buy/sell signals generated from Gen-AI technique can outperform a passive index by 10-30%.

Additionally, Lezhi et al. [68] showcased how multimodal Gen-AI techniques could be used for fundamental investment research. Through fine-tuning methods applied to a base model (Llama2), they developed an AI (Artificial Intelligence) agent that can assist investors in tasks such as understanding market conditions, generating investment ideas, and formatting results with stock recommendations.

C. Summarization and Assistive Applications

Gen-AI techniques have also been applied in text summarization and assistive applications. In particular, Xiao et al. [69] observed a clear preference among human evaluators for LLM-generated summaries over human-written summaries. Gen-AI techniques like SaulLM-7B [70] based on mistral 7B architecture, are designed for legal text generation and comprehension.

Further, JP Morgan uses a Gen-AI toolkit, designed to serve as a 'research analyst', aiding in various tasks that enhance productivity and decision-making within the firm [71]. Furthermore, companies like Kudos and Max-Rewards understand credit card offerings and combine them with user's spends to suggest the best credit cards for maximising rewards. Morgan Stanley's COIN (Contract Intelligence) uses AI to review legal documents and extract relevant information.

Leading CRM companies such has Salesforce [72], Hubspot [73], Zendesk [74] offer AI based chatbots for better customer experience. Further, Intuit has integrated, a Gen-AI powered financial assistant called intuit assist into TurboTax [75], that uses a combination of the company's prior tax preparation experience, data and documents from the filer, and current tax code for tax filing.

D. Risk Monitoring

Mastercard [76] discussed how Gen-AI techniques could help financial institutions improve their fraud detection rates by 20%, on average. Further, Cao et al. [77] explored a new framework that leverages Gen-AI to analyze and predict financial risks. It uniquely combines different types of financial data, including textual and vocal information from Earnings Conference Calls (ECCs), market-related time series data, and contextual news data and showcases the critical role of LLMs in financial risk assessment and opens new avenues for their application in this field.

VI. Conclusion

Compared to the traditional machine learning and data mining approaches, Gen-AI techniques are able to improve understanding of context and nuances in language modeling, translation between languages, handle large volumes of data, provide fast and low-latency responses and can be fine-tuned for various tasks and domains. In this manuscript we have provided a brief overview of the application of Gen-AI techniques within the finance ecosystem.

Further, we discussed the various opportunities and challenges associated with the widespread adoption of Gen-AI techniques in the finance domain. We also discuss techniques which can be used to develop Gen-AI models by overcoming challenges stemming from critical issues such as hallucination, data-quality and computational costs. Lastly, we provide application areas in Finance where these Gen-AI models are or can be used.

To the best of our knowledge, this is the first work which comprehensively summarizes the usage of Gen-AI techniques in the finance domain. This summarization would not only provide a clear overview of the areas where substantial work has been completed but also pinpoint areas where future efforts can be prioritized, potentially accelerating the development and adoption of Gen-AI techniques in finance. Moreover, this analysis could serve as a valuable bridge between finance and other domains, facilitating the cross-pollination of innovative ideas and best practices, ultimately benefiting a wider range of industries and applications.

References

[1] Lior Rokach and Oded Maimon. Clustering methods. In *Data mining and knowledge discovery handbook*, pages 321–352. 2005.

[2] John Aitchison and Ian Robert Dunsmore. *Statistical prediction analysis*. 1975.

[3] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

[4] James D. Hamilton. *Time series analysis*. Princeton university press, 2020.

[5] Y. Kashyap and A. Sinha. Llm is all you need: How do llms perform on prediction and classification using historical data. *International Journal For Multidisciplinary Research*, 6(3), Jul 2024.

[6] J. et al. Yang. Harnessing the power of LLMs in practice: A survey on chatgpt and beyond. *ACM Trans. Knowl. Discov. Data*, 18(6):1–32, Jul 2024. doi: 10.1145/3649506.

[7] Y. et al. Xia. Understanding the performance and estimating the cost of llm fine-tuning. *arXiv:2408.04693*, 2024.

[8] Ashish et al. Vaswani. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[9] M. Tan and Q. Le. Efficientnetv2: Smaller models and faster training. In *Proc. Int. Conf. Mach. Learn.*, pages 10096–10106, 2021.

[10] W. et al. Sun. CEBench: A benchmarking toolkit for the cost-effectiveness of llm pipelines. *arXiv:2407.12797*, 2024.

[11] C. Milana and A. Ashta. Artificial intelligence techniques in finance and financial markets: A survey of the literature. *Strategic Change*, 30(3):189–209, 2021.

[12] G. Shabsigh and E. Boukherouaa. *Generative Artificial Intelligence in Finance: Risk Considerations*. International Monetary Fund, 2023.

[13] B. Singh. Generative artificial intelligence: Prospects for banking industry. *International Journal of Research in Engineering, Science and Management*, 7(3):83–86, 2024.

[14] H. et al. Zhao. Revolutionizing finance with llms: An overview of applications and insights. *arXiv:2401.11641 [cs.CL]*, 2024.

[15] X. et al. Cao. Empowering financial futures: Large language models in the modern financial landscape. *EAI Endorsed Transactions on AI and Robotics*, 3, 2024.

[16] A. Goldberg. Ai in finance: Leveraging large language models for enhanced decision-making and risk management. *Social Science Journal for Advanced Research*, 4(4):33–40, 2024.

[17] Libo et al. Qin. End-to-end task-oriented dialogue: A survey of tasks, methods, and future directions. *arXiv preprint arXiv:2311.09008*, 2023.

[18] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[19] The economic potential of generative ai: The next productivity frontier.

[20] Klarna ai assistant handles two-thirds of customer service chats in its first month. Accessed: 2024-10-27.

[21] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

[22] T. Schick, J. Dwivedi-Yu, R. Dess‘ı, R. Raileanu, M. Lomeli, L. Zettlemoyer, N. Cancedda, and T. Scialom. Toolformer: Language models can teach themselves to use tools. 2023.

[23] S. et al. Zhang. Llasa: Large language and e-commerce shopping assistant. *arXiv:2408.02006*, 2024.

[24] Shamima Ahmed, Muneer M. Alshater, Anis El Ammari, and Helmi Hammami. Artificial intelligence and machine learning in finance: A bibliometric review. *Research in International Business and Finance*, 61:101646, 2022. ISSN 0275-5319. doi: https://doi.org/10.1016/j.ribaf.2022.101646.

[25] Jin K Kim, Michael Chua, Mandy Rickard, and Armando Lorenzo. Chatgpt and large language model (llm) chatbots: The current state of acceptability and a proposal for guidelines on utilization in academic medicine. *Journal of Pediatric Urology*, 19(5):598–604, 2023.

[26] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *arXiv:2201.11903*, 2022.

[27] C. et al. Maple. The ai revolution: opportunities and challenges for the finance sector. *arXiv preprint arXiv:2308.16538*, 2023.

[28] S. et al. Wu. Bloomberggpt: A genai model for stock prediction and trading. *arXiv:2303.17564*, 2023.

[29] Kruse et al. Artificial intelligence for the financial services industry: What challenges organizations to succeed, 2019.

[30] S. A. et al. Assefa. Generating synthetic data in finance: opportunities, challenges and pitfalls. In *Proc. First ACM*

Int. Conf. AI in Finance, pages 1–8, 2020.

[31] Mikael Ljung. Synthetic data generation for the financial industry using generative adversarial networks, 2021.

[32] J. et al. Devlin. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*, 2019.

[33] Y. et al. Li. Large language models in finance: A survey. In *Proceedings of the fourth ACM international conference on AI in finance*, pages 374–382, 2023.

[34] Z. et al. Liu. Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*, pages 4513–4519, 2021.

[35] Hu et al. LORA: LOW-RANK ADAPTATION OF LARGE LANGUAGE MODELS. *arXiv:2106.09685*, 2021.

[36] Y. et al. Xia. Understanding the performance and estimating the cost of llm fine-tuning. *arXiv:2408.04693 [cs.CL]*, 2024.

[37] D. Hancock and D. B Humphrey. Payment transactions, instruments, and systems: A survey. *Journal of Banking & Finance*, 21(11):1573–1624, 1997.

[38] C. et al. Bryce. Trends in large language models: Actors, applications, and impact on cybersecurity.

[39] V. Srivastava. BAI-Arg LLM at the finllm challenge task: Earn while you argue-financial argument identification. In *Proc. Eighth Financial Technology and Natural Language Processing and the 1st Agent AI for Scenario Planning*, pages 165–173, 2024.

[40] K. et al. Lakkaraju. LLMs for financial advisement: A fairness and efficacy study in personal decision making. In *Proc. Fourth ACM Int. Conf. AI in Finance (ICAIF '23)*, pages 100–107, Brooklyn, NY, USA, 2023. doi: 10.1145/3604237.3626867.

[41] S. Glavina. AI IN FINANCIAL INDUSTRY: ETHIC ISSUES, 2024.

[42] K. Huang, J. Huang, and D. Catteddu. Genai data security. In *Generative AI Security: Theories and Practices*, pages 133–162. Springer, 2024.

[43] D. et al. Zhang. Right to be forgotten in the era of large language models: Implications, challenges, and solutions. *arXiv:2307.03941*, 2023.

[44] J. Rosen. The right to be forgotten. *Stan. L. Rev. Online*, 64:88, 2011.

[45] Ggaliwango et al. Marvin. Prompt engineering in large language models. In *International conference on data intelligence and cognitive informatics*, Singapore, 2023. Springer Nature Singapore.

[46] Wei et al. Wang. A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–37, 2019.

[47] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[48] Elhage et al. A mathematical framework for transformer circuits. Transformer Circuits Thread, 2021.

[49] Olsson et al. In-context learning and induction heads. Transformer Circuits Thread, 2022.

[50] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023.

[51] Rabeeh Karimi Mahabadi, Sebastian Ruder, Mostafa Dehghani, and James Henderson. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. *arXiv preprint arXiv:2106.04489*, 2021.

[52] Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965, 2022.

[53] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *CoRR*, abs/2104.08691, 2021.

[54] Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*, 2024.

[55] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*, 2020.

[56] A. Patel, S. Bhattamishra, and N. Goyal. Are nlp models really able to solve simple math word problems? *CoRR*, abs/2103.07191, 2021.

[57] Markus Nagel, Marios Fournarakis, Rana Ali Amjad, Yelysei Bondarenko, Mart van Baalen, and Tijmen Blankevoort. A white paper on neural network quantization. *CoRR*, abs/2106.08295, 2021.

[58] Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W. Mahoney, and Kurt Keutzer. A survey of quantization methods for efficient neural network inference. *CoRR*, abs/2103.13630, 2021.

[59] Yoshua Bengio. Estimating or propagating gradients through stochastic neurons. *CoRR*, abs/1305.2982, 2013.

[60] Zhiyu et al. Chen. Finqa: A dataset of numerical reasoning over financial data. *arXiv preprint arXiv:2109.00122*, 2021.

[61] Zhiyu et al. Chen. Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering. *arXiv preprint arXiv:2210.03849*, 2022.

[62] Xianzhi et al. Li. Are chatgpt and gpt-4 general-purpose solvers for financial text analytics? a study on several typical tasks. *arXiv preprint arXiv:2305.05862*, 2023.

[63] Fintool team. Fintool. https://fintool.com/.

[64] R. Graciano. Intuit credit karma scales genai-first experiences. *Credit Karma*, July 2024.

[65] D. Mai. Stockgpt: A genai model for stock prediction and trading. *arXiv preprint arXiv:2404.05101*, 2024.

[66] A. Lopez-Lira and Y. Tang. Can chatgpt forecast stock

price movements? return predictability and large language models. *SSRN Electronic Journal*, 2023.

[67] Georgios et al. Fatouros. Can large language models beat wall street? unveiling the potential of ai in stock selection. *arXiv preprint arXiv:2401.03737*, 2024.

[68] L. et al. Li. Multimodal gen-ai for fundamental investment research. *arXiv preprint arXiv:2401.06164*, 2023.

[69] Xiao Pu, Mingqi Gao, and Xiaojun Wan. Summarization is (almost) dead. *arXiv preprint arXiv:2309.09558*, 2023.

[70] P. et al. Colombo. SaulLM-7B: A pioneering large language model for law. *arXiv preprint arXiv:2403.03883*, 2024.

[71] Jpmorgan chase leads ai revolution in finance with launch of llm suite. *Forbes*, Jul 2024.

[72] Salesforce. Einstein ai assistant.

[73] Hubspot. Hubspot chatbot builder.

[74] Zendesk. Zendesk chatbot builder.

[75] Lisa Greene-Lewis. Transformative tax preparation: Gen-ai powered intuit assist, September 2023.

[76] R. Brown. Mastercard launches gpt-like ai model to help banks detect fraud. *CNBC*, Feb 2024.

[77] Yupeng Cao, Zhi Chen, Qingyun Pei, Fabrizio Dimino, Lorenzo Ausiello, Prashant Kumar, K. P. Subbalakshmi, and Papa Momar Ndiaye. Risklabs: Predicting financial risk using large language model based on multi-sources data. *arXiv preprint arXiv:2404.07452*, 2024.