Generative Artificial Intelligence in Finance: Risk Considerations

Ghiath Shabsigh and El Bachir Boukherouaa

NOTE/2023/006

FINTECH NOTE

Generative Artificial Intelligence in Finance: Risk Considerations

Prepared by Ghiath Shabsigh and El Bachir Boukherouaa

August 2023

Generative Artificial Intelligence in Finance: Risk Considerations
NOTE/2023/006

Ghiath Shabsigh and El Bachir Boukherouaa

Cataloging-in-Publication Data
IMF Library

Names: Shabsigh, Ghiath, author. | Boukherouaa, El Bachir, author. | International Monetary Fund, publisher.
Title: Generative artificial intelligence in finance : risk considerations / Prepared by Ghiath Shabsigh and El Bachir Boukherouaa.
Other titles: Risk considerations. | Fintech notes.
Description: Washington, DC : International Monetary Fund, 2023. | Aug. 2023. | Note 2023/006. | Includes bibliographical references.
Identifiers: ISBN:
9798400251092 (paper)
9798400250569 (ePub)
9798400251443 (webPDF)
Subjects: LCSH: Artificial intelligence—Financia applications. | Finance—Technological innovations.
Classification: LCC HG4515.5 S53 2023

RECOMMENDED CITATION: Shabsign, Ghiath, and El Bachir Boukherouaa. 2023. “Generative Artificial Intelligence in Finance: Risk Considerations.” IMF Fintech Note 2023/006, International Monetary Fund, Washington, DC.

Publication orders may be placed online, by fax, or through the mail:

International Monetary Fund, Publications Services
P.O. Box 92780, Washington, DC 20090, USA
Tel.: (202) 623-7430 Fax: (202) 623-7201
E-mail: publications@imf.org
bookstore.IMF.org
elibrary.IMF.org

Contents

Abbreviations **2**
Introduction **3**
Risk Considerations **4**
Conclusion **11**

Appendix 1. Comparisons of Main Large Language Models **13**

Appendix 2. Generative AI: Stylized Architecture **15**

References **19**

BOXES

Box 1. Sample Generative AI Applications in the Financial Sector 4
Box 2. Synthetic Data in AI 8

FIGURES

Figure 1. Months to Reach 100 Million Users 3
Appendix Figure 2.1. The Transformer Model's Architecture 16

TABLE

Appendix Table 1.1. Comparisons of Main Large Language Models 13

Abbreviations

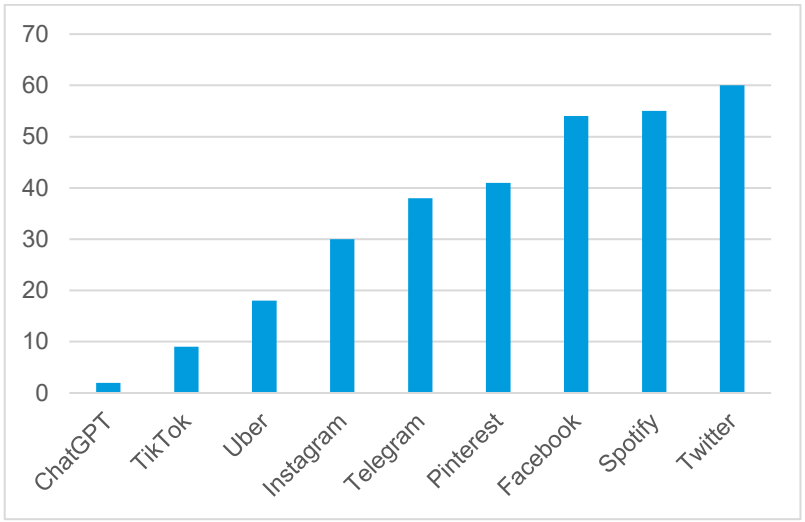
AI	Artificial intelligence
AI/ML	AI–machine learning
GenAI	Generative artificial intelligence
LLM	Large language model
NLP	Natural language processing
RNNs	Recurrent neural networks

Introduction

Artificial intelligence (AI) has enormous transformative power and holds profound implications for the world's societies and economies. AI is playing an increasingly important role in shaping economic and financial sector developments and is seen as an engine of productivity and economic growth through efficiency, improved decision-making processes, and the creation of new products and industries.[1] AI also is rapidly changing the financial sector landscape by reshaping the nature of financial intermediation, risk management, compliance, and prudential oversight.

Upon its launch on November 30, 2022, Chat Generative Pre-Trained Transformer (ChatGPT) triggered massive global reaction. Remarkably, within a span of two months, the platform gained more than 100 million active users across the globe, a rate much faster than that of other platform innovations (Figure 1). The users represented a broad spectrum (for example, industries, academia, legal firms, and publishing houses), all of which have started leveraging the technology's capabilities. By March 2023, several competitors had introduced their own iterations (see Appendix 1) of what are now commonly referred to as generative AI systems (GenAI).

Figure 1. Months to Reach 100 Million Users



Source: IMF staff calculations based on data from company websites.

GenAI is a significant leap forward in AI technology. GenAI is a specific subset of AI–machine learning (AI/ML) technologies, distinguished by their ability to create new content. At the heart of GenAI are large language models (LLMs), which are neural network–based models trained on massive amounts of data, including text and documents, and capable of producing understandable and meaningful text or

[1] For a broader discussion of AI economics, see, for example, Agrawal, Gans, and Goldfarb (2018) and Acemoglu and Restrepo (2019).

human languages (Appendix 2). LLMs enable a wide range of applications across various domains with significant implications for the global economy and financial sector.

GenAI will accelerate AI adoption in the financial sector. Competitive pressures have fueled rapid adoption of AI/ML in the financial sector in recent years by facilitating gains in efficiency and cost savings, reshaping client interfaces, enhancing forecasting accuracy, and improving risk management and compliance. GenAI could also deliver to cybersecurity benefits ranging from implementing predictive models for faster threat detection to improved incident response. Financial service providers have been quick to explore the capabilities of GenAI and how it can be adapted to a broad range of applications (Box 1). GenAI's ability to process very large and diverse data sets and to generate content in accessible and easily usable formats (including conversational) is proving very useful in enhancing efficiency and improving customer experience, risk mitigation, and compliance reporting for financial providers. However, the deployment of GenAI in the financial sector has its own risks that need to be fully understood and mitigated by the industry and prudential oversight authorities.

Box 1. Sample Generative AI Applications in the Financial Sector

- Capital One and JPMorgan Chase have leveraged GenAI to augment their AI-powered fraud and suspicious activity detection system. This effort seems to have resulted in a significant reduction in false positives, a better detection rate, reduced costs, and improved customer satisfaction.
- Morgan Stanley Wealth Management will use OpenAI's technology to leverage its own vast data sources to assist financial advisors with insights into companies, sectors, asset classes, capital markets, and regions around the world.
- Wells Fargo is building capabilities for automating document processing, including providing summary reports, and scaling up its virtual assistant chatbots.
- Goldman Sachs and Citadel are considering GenAI applications for internal software development and information analysis.

This note builds on the 2021 IMF Paper that assessed AI/ML risks for the financial sector by examining the characteristics that differentiate GenAI from AI/ML and the new risks that unique aspects may raise (Boukherouaa and Shabsigh 2021). The wide-ranging appeal of GenAI technology combined with its new complex risks will have broad systemic implications for the financial sector. Rather than a technical discussion of GenAI, this note seeks to explore the potential risks to the financial sector from this technology based on its current technical characteristics.

Risk Considerations

The deployment of AI applications in the financial sector is raising several concerns about the risks inherent in the technology. These concerns include embedded bias and privacy shortcomings, opaqueness about how outcomes are generated, robustness issues, cybersecurity, and AI's impact on broader financial stability. Concerns about risks inherent in GenAI applications are broadly similar to

those about AI/ML but with important variations that would need to be considered carefully by the industry and prudential oversight authorities, as detailed below. Furthermore, the distribution of risks between public and private GenAI applications may vary and the risks are likely to be better managed in the latter.

The launch of ChatGPT has generated fears about the potential risks that GenAI poses. Several major financial institutions have reportedly barred employees from using ChatGPT.[2] In April 2023 Italy temporarily banned ChatGPT over concerns regarding potential violations of the European Union's General Data Protection Regulations. The US Consumer Financial Protection Bureau is closely examining and monitoring GenAI's potential risks to the financial sector, including from bias or misleading information. Calls in the European parliament have been made to augment the proposed "European AI Act" with specific provisions for GenAI.

This note explores the risks posed by using GenAI systems in the financial sector. These risks include those inherent in the technology (data privacy and embedded bias), those related to its performance (robustness, synthetic data, and explainability), new cybersecurity threats posed by GenAI, and broader risks to financial stability.

Data Privacy

AI/ML systems raise several well-known privacy concerns, and they must be addressed when AI/ML is used in the highly regulated financial sector. They include, among others, data leakages from the training data sets,[3] the capacity to unmask anonymized data through inferences,[4] and AI/ML "remembering" information about individuals in the training data set after the data are used and discarded; further, AI/ML's output may leak sensitive data directly or by inference. These concerns are at the heart of ongoing efforts to improve AI/ML privacy and update the legal and regulatory framework that requires AI/ML systems and related data sources to adhere to enhanced privacy standards. GenAI raises privacy issues that are similar to those of AI/ML, but it also raises new, unique concerns.

Publicly available GenAI systems pose significant privacy challenges for financial institutions wishing to incorporate their capabilities into their operations. By automatically "opting in" every user, these GenAI systems continuously use inputs from users for training and for fine-tuning their responses.[5] This automation thus raises the possibility that sensitive financial data and personal information provided by financial institutions' staff in their engagement with the GenAI could leak out. Several GenAI systems often explicitly state that they cannot ensure the security and confidentiality of the information and data provided by users.

[2] These include Goldman Sachs, JPMorgan Chase, Citigroup, Bank of America, Deutsche Bank, and Wells Fargo. See Retail Banker International 2023.

[3] Data leakages could very well expand beyond private and personal data to proprietary and confidential financial sector data.

[4] This refers to AI/ML's capacity to deduce identities from behavioral patterns.

[5] The "opting out" choice for user data collection and use needs to be explicitly exercised. However, opting out seems to limit, although it's unclear to what extent, GenAI responses and thus possibly diminishes the technology's utility.

Enterprise-level GenAI systems are being developed, in part, to address privacy concerns associated with public GenAI, but some privacy concerns will likely persist. In principle, these enterprise-level GenAI could improve data security for financial institutions. However, residual privacy concerns remain. They relate to the nature of GenAI's capabilities to process a broad range of data formats, including scraping information from internet and online platforms (for example, social media). These data make GenAI a valuable tool for financial institutions to use for applications like fraud detection and credit assessment. However, this capability comes with the risk of unintentionally collecting and using personal information that otherwise may have needed explicit consent.

Embedded Bias

An important challenge for AI systems is embedded bias—particularly in a highly regulated and sensitive sector like financial services. Embedded bias could be defined as computer systems that systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others (Friedman and Nissenbaum 1996). Bias could emerge if the data used to train the system are incomplete or unrepresentative, or the data are underpinned by prevailing societal prejudices. Bias could also arise in the AI algorithm if its design is influenced by human biases. In the financial sector, which is increasingly dependent on AI-supported decisions, embedded bias could lead to, among other things, unethical practices, financial exclusion, and damaged public trust.

GenAI could aggravate the embedded bias problem. GenAI models are trained on a broad set of online textual and other data formats that inherently carry in them real-life human biases. In the case of AI/ML, operators try to mitigate the embedded bias through the selection process of the data used to train the AI/ML. This process, however, could be far more complicated for GenAI, given the breadth and diversity of the training data. Furthermore, bias could arise from the process and algorithm used in generating GenAI responses. Unlike AI/ML, which uses training data for predictions, GenAI uses its training data to create textual answer, that is "new content", based on accuracy probability of each part of the answer. This answer, in turn, is influenced by prompts directed to the GenAI; the prompts themselves could also carry in them human biases.[6]

GenAI could be susceptible to bias generated by search engine optimization (SEO) tools (see, for example, Atreides 2023). To improve their visibility in internet search engines (for example, Google, Bing, and others), websites use SEO techniques. SEO is primarily used, at present, for marketing products and services or disseminating information. As the use of GenAI applications spreads, SEO tools will very likely be geared toward influencing the training of GenAI models—possibly skewing the models output and introducing new layers of biased data that could be difficult to detect.

The data bias problem in GenAI could complicate its adoption and use in financial services. GenAI could offer a quick and low-cost way for financial institutions to profile their clients, including for risk management, and to screen transactions with a view to identifying the ones that are suspicious. However, the potential of overreliance on GenAI-generated profiles, without appropriate

[6] Preliminary studies show that over time the biases introduced by GenAI may become even more perpetuated and worse than reality; see Nicoletti and Bass 2023.

safeguards, could lead to inaccurate or discriminatory client assessments. Appropriate human judgment will need to complement GenAI-based transaction monitoring models. Furthermore, GenAI-based chatbots constitute a particularly sensitive issue when the system is used to address inquiries and complaints by clients, who may not realize they are dealing with an automated system. Such systems could misdirect certain client segments, reflecting embedded biases. Nevertheless, the use of chatbots does not excuse a financial institution from its legal and regulatory obligations (see, for example, US Consumer Financial Protection Bureau 2023).

Robustness

Robust AI performance in the financial system is rapidly becoming an important issue for safeguarding financial stability and integrity and, ultimately, maintaining public trust. Robustness covers issues related to the accuracy of AI models' output, particularly in a changing environment. It also covers governance of the development and operation of AI systems to safeguard against unethical use, including exclusionary, biased, and harmful outcomes.[7]

Given the predictive nature of AI/ML algorithms, a key challenge is their ability to minimize false signals during periods of structural shifts. AI/ML models seem to perform well in a relatively stable data environment that produces reliable signals, enabling the models to incorporate evolving data trends without significant loss in prediction accuracy. However, AI/ML models face a more challenging task when previously reliable signals become unreliable or when behavioral correlations shift significantly, leading to a loss in prediction accuracy.

GenAI models face different challenges to performance robustness, reflecting the nature of GenAI's data environment and decision-making process. GenAI's ability to generate new content based on training data comes with the risk that GenAI models could produce wrong but plausible-sounding answers or output and then defend those responses confidently—a phenomenon broadly referred to as "hallucination." The problem is even more acute in conversational GenAI, which could amplify instances of hallucinations (Dziri and others 2022). Although what causes the phenomenon is not yet fully understood, several factors have been proposed, such as information misalignment or divergence between reference and source data, which is a possibility in large data sets, and how the model is developed and trained (Parikh and others 2020).

There are ongoing efforts at present to address GenAI hallucination, but they are narrowly focused on specific tasks (for example, abstractive summarization) rather than on addressing the problem from a broader perspective (Ji and others 2023). Efforts to develop enterprise-level GenAI could help minimize the problem by providing more focused, better quality, and more transparent training data. The hallucination risk, however, will likely remain a concern in the foreseeable future.

In the context of financial services, GenAI hallucination is a significant risk on multiple levels. It undermines GenAI robustness and raises financial safety and consumer protection concerns.

[7] This section does not discuss robustness issues with respect to GenAI governance, as they are broadly similar to those related to AI/ML (see Boukherouaa and others 2021).

For example, GenAI-generated risk assessment reports based on market sentiments, or customer profile reports from online sources, could be wrong, and this inaccuracy has negative implications for risk-taking and management. Financial services offered to customers through GenAI-supported conversational bots could give inappropriate advice or offer the wrong product to undiscerning clients. These and similar outcomes will expose the financial system to significant risks and erode public trust in AI systems and the financial institutions using them.

Synthetic Data

The use of synthetic data in the context of AI systems has accelerated in recent years. Synthetic data are algorithm-created with a statistical distribution that mimics real data via deep learning model simulation. Synthetic data are used primarily to train AI/ML and for testing model robustness (Box 2). Synthetic data have emerged as a viable alternative to real data primarily because of their ability to alleviate privacy and confidentiality concerns—coupled with their cost-effectiveness. Nevertheless, the use of synthetic data poses several challenges, notably issues pertaining to data quality along with the potential for replication of inherent real-world biases and gaps in the generated data sets. Major technology companies have turned to synthetic data to address a spectrum of operational challenges and objectives. Apple, for instance, employs synthetic data to enhance Siri's voice recognition capabilities, while Tesla deploys synthetic data in the simulation of myriad driving conditions. Firms in the retail sector are adopting synthetic data to emulate consumer behavior patterns, thereby gaining actionable insights.

Box 2. Synthetic Data in AI

The spread of the use of synthetic data is driven primarily by regulatory necessity and practical business needs. Concerns about data privacy in the context of AI/ML training, particularly in highly regulated sectors like financial and health services, make the use of synthetic data, which cannot be attributed to any person or group, an attractive solution. Synthetic data also offer the opportunity to mitigate imbalances and biases in real data and help build more robust, explainable models that better meet regulatory requirements (Papenbrock and Ebert 2022). In addition, given that the data ownership environment at present is highly lopsided to the benefit of established technology and industry incumbents, synthetic data provide a cost-efficient training data alternative to those businesses that lack significant access to proprietary real data.

GenAI could significantly expand the horizons for the use of synthetic data in the financial sector. GenAI is intrinsically geared toward generating new content and using more diverse sets of data sources, it can be used to code synthetic data–generator algorithms, and it better captures the complexity of real-world events. These properties are proving attractive to financial institutions, as they can customize their AI training to specific functions (for example, fraud detection), product development and delivery, and compliance reporting.

It is not clear, however, to what extent GenAI could impart some of its risks (for example, bias, accuracy) to the generated synthetic data. If so, this ability will undermine the quality of the synthetic data and their usefulness for training AI/ML systems. The attractiveness of GenAI for generating synthetic data coupled with the complexity of how the data are generated could potentially blind financial institutions to the potential risks the training data are embedding into their operations.

Explainability

Financial institutions are required to be able to explain their decisions and actions, internally and to external stakeholders, including prudential supervisors. These decisions involve developing and marketing products, managing risks, fulfilling regulatory requirements (such as obligations related to anti–money laundering and combating the financing of terrorism), and engaging consumers. Being able to explain financial decisions is at the core of sound financial systems.

But ensuring the explainability of decisions and actions taken as an outcome of AI algorithms is a complex and multifaceted issue. AI algorithms have dense architecture that relies on numerous parameters and are often an ensemble of interacting models, and whose input signals might not be easily identifiable or even known. Furthermore, there is a general trade-off between model accuracy and flexibility,[8] and its explainability.

The emergence of GenAI has exacerbated the AI explainability problem. The breadth and diversity of the data used by GenAI— which are at the core of its utility—make it exceedingly difficult at present to map GenAI's output to the data, including in the extreme case of hallucination. Furthermore, GenAI's architecture and decision-making process contribute greatly to the opaqueness of GenAI's output process. GenAI algorithms runs on multiple neural network layers and uses numerous parameters to calculate the probabilities of each part of its answers.

GenAI explainability will be a challenge for the financial sector's GenAI adoption. Research is ongoing to develop solutions that could improve GenAI explainability (see, for example, Ullah and others 2020). Indeed, because of the ingestion of the massive data and the complexity of the algorithms and the architecture of LLM, explainability or interpretability in GenAI systems continues to be a challenge for the research community. Some techniques have been proposed recently to provide insight on the outcome of those models, but the result remains unsatisfactory. This problem persists; thus, the adoption of those models in the financial sector requires more scrutiny. GenAI output does not consist of decisions but of texts. Accordingly, the proper domain of GenAI is recommendations, advice, or analysis, where human actors should make decisions and assume the responsibility for them. The nuance is that financial institutions need to understand the reasons for their actions, and where these actions are based on outputs generated by GenAI, these institutions should be able to understand the generative process and its limitations.

[8] Here, the term *flexibility* refers to algorithm's capacity to approximate different functions and is directly related to the number of the model's parameters.

Cybersecurity

GenAI poses significant new challenges to the cybersecurity landscape. This emerging technology could be exploited to generate more sophisticated phishing messages and emails or to present opportunities for malicious actors to impersonate individuals or organizations, leading to increased identity theft or fraud. The proliferation of deepfakes, resulting in more realistic videos, audios, or images, could inflict serious damage on both organizations and individuals.

GenAI models could be vulnerable to data poisoning and input attacks (see Boukherouaa and others 2021). Data poisoning attacks attempt to influence AI models at the training stage by adding special elements to the training data set; the effort seeks to undermine training accuracy or to hide malicious actions that wait for special inputs. Input attacks are similar, but they attempt to influence the AI models during operation. GenAI could be susceptible to similar data manipulation attacks. Tools, like SEO or GenAI-generated content, could potentially be used to manipulate the GenAI data environment for malicious purposes. While at present this risk may not be material because current GenAI models are trained and operate on pre-2021 internet scraped data, the situation could quickly change as more people are aware of GenAI capabilities and rapid adoption. Moreover, enterprise-level GenAI applications could be particularly vulnerable, as they use more focused data sets that could be targeted by purpose-built cyberhacking tools.

Current GenAI models are increasingly subject to successful "jailbreaking" attacks (see, for example, ADVERSA 2023). These attacks rely on developing sets of carefully designed prompts (word sequences or sentences) to bypass GenAI's rules and filters or even insert malicious data or instructions (the latter is sometimes referred to as "prompt injection attack"). These attacks could corrupt GenAI operations or siphon out sensitive data.

Given that GenAI technology is a relatively new phenomenon, the full scale of its vulnerability to cyberattacks is yet to be comprehensively understood. Nevertheless, early signs indicate potentially substantial issues that warrant careful contemplation, especially when decision makers are considering large-scale adoption of the technology in sensitive and heavily regulated sectors such as finance—and particularly in the case of enterprise-level GenAI systems.

Financial Stability

As highlighted in the IMF 2021 paper, AI/ML could potentially bring about new sources and transmission channels of systemic risks. In particular, the widespread use of AI/ML could drive greater homogeneity in risk assessments and credit decisions in the financial sector, as well as out-of-sample risk that, coupled with rising interconnectedness, could create the conditions for a buildup of systemic risks. AI/ML may also automate and accelerate the procyclicality of financial conditions through, for example, automating AI/ML's risk assessments and credit underwriting decisions that are inherently procyclical. In the case of a tail risk event, AI/ML could quickly amplify and spread the shock throughout the financial system and complicate the effectiveness of the policy response.

GenAI would likely bring about systemic risks similar to those of AI/ML, but it would also bring its own concerns. These concerns could be exacerbated by the ease and cost-effectiveness with which GenAI reports can be generated and the lack of effective regulatory regime. This environment could increase the temptation for excessive reliance on GenAI, which, in turn, could increase contagion risk and build systemic risks in the financial sector.

- Decisions made by financial institutions based on GenAI-generated economic, market, or risk reports could be susceptible to herd mentality bias and mispricing risk if these reports reflect public sentiments captured from the data sets used by the GenAI system, particularly at times of market euphoria.[9]

- GenAI hallucination is a concern that could become systemically important if the misleading information spreads in the financial system, aided by the concentration of GenAI service providers, and because of the challenges in interpreting and identifying sources and counterparties.[10]

- GenAI could generate solvency and liquidity risks if AI-driven trades take higher-credit or market risks to maximize profit if the models are not trained about risk management properly. The herding behavior of GenAI investment advisors could affect market liquidity, and rumors propagated by GenAI could trigger bank runs.

- Cybersecurity of GenAI, including potential susceptibility to data-manipulation attacks, is a particular concern given the potential of GenAI to generate false and malicious content. Such content could create public panic, which in the case of financial services could result, for example, in bank runs.

Conclusion

GenAI technologies hold great promise for financial sector applications but should be approached with caution. GenAI could drive significant efficiency, improve customer experience, and strengthen risk management and compliance. However, the *intrinsic* risks in GenAI could pose *material* risks for financial sector reputation and soundness—and, ultimately, could undermine public trust. Enterprise-level GenAI applications could help mitigate some of the risks inherent in public GAIs, but this option may not be cost efficient for smaller financial institutions.

[9] GenAI applications could contribute to liquidity risk if their algorithms inadvertently promote among market participants herd behavior resulting in simultaneous buying or selling decisions; large-scale market dislocations could result.

[10] Financial institutions' reliance on GenAI technologies from a small number of providers could result in concentration risk and lead to making those providers vulnerable to various operational risks or disruptions. The high dependence of financial business and the technological concentration could create a “too-big-to-fail” problem.

Regulatory policy will evolve over time to help guide the use of GenAI applications by financial institutions, but interim actions are needed. GenAI use needs close human supervision commensurate with the risks that could materialize from employing the technology in financial institutions' operations (for example, the use of AI for analysis or recommendations vs. the implementation of AI systems that have the capacity to make and execute decisions). Prudential oversight authorities should strengthen their institutional capacity and intensify their monitoring and surveillance of the evolution of the technology, paying close attention to how it is applied in the financial sector. To do so, they should improve communication with public and private sector stakeholders as well as collaborate with jurisdictions at the regional and international levels.

Appendix 1. Comparisons of Main Large Language Models

Appendix Table 1.1. Comparisons of Main Language Models

LLM	Company	Parameters	Release Date	Performance	Main Application
BERT	Google AI	340 million	October 2018	State of the art in a variety of NLP tasks, including question answering and natural language inference. GLUE score: 86.5	NLP, including question answering and natural language inference
Turing-NLG	Microsoft	17 billion	February 2020	Achieved state-of-the-art results using transformer architecture. In a recent benchmark study, Turing-NLG outperformed other LLMs on several tasks including GPT. GLUE score: 92.8	NLP, including question answering, natural language inference, and text summarization
Megatron-Turing NLG	Google AI	530 billion	October 2021	Achieved state-of-the-art results on the SuperGLUE benchmark for natural language understanding. GLUE score: 92.6	NLP, including natural language understanding and natural language generation
LaMDA	Google AI	137 billion	May 2022	Achieved state-of-the-art results on the C4 benchmark for commonsense reasoning. GLUE score: 93.4	Conversational AI, including question answering and generating different creative formats of text content
Blender	Blender Institute, Netherlands	137 billion	May 2022	Achieved state-of-the-art results on the C4 benchmark for commonsense reasoning. GLUE score: 92.9	NLP, including question answering, natural language inference, and creative writing
Jurassic-1 Jumbo	Google AI	1.75 trillion	June 2022	Achieved state-of-the-art results on the GLUE benchmark for natural language understanding. GLUE score: 94	NLP, including natural language understanding and natural language generation
WuDao 2.0	Beijing Academy of Artificial Intelligence	1.75 trillion	June 2022	Achieved state-of-the-art results on the GLUE benchmark for natural language understanding. GLUE score: 94.2	NLP, including machine translation and question answering

GPT-3	OpenAI	175 billion	November 2022	State of the art in a variety of natural language processing (NLP) tasks, including machine translation, text summarization, and question answering. GLUE[11] score: 80.3	NLP, including machine translation, text summarization, and question answering
GPT-4	OpenAI	175 billion	March 2023	Better than GPT-3 on complex tasks, though slower and more expensive. GLUE score: 93.3	Similar to GPT-3
Claude	Anthropic	137 billion	March 2023	Achieved state-of-the-art results on the GLUE benchmark for natural language understanding. GLUE score: 92.2	Good on both broad and complex tasks, more balanced than GPT-4
Pi	Inflection	137 billion	June 2023	Achieved state-of-the-art results on the GLUE benchmark for natural language understanding. GLUE score: 92.8	Excellent on both broad and complex tasks, comparable with GPT-4

[11] The GLUE (General Language Understanding Evaluation) benchmark is a set of nine task databases designed to evaluate and score a model's language understanding.

Appendix 2. Generative AI: Stylized Architecture

LLMs are machine learning models that are good at understanding questions or requests and generating human language. Those models operate by ingesting vast quantities of data, for training purposes, to discern statistical patterns, including the relationships between words and the contextual significance of each word within a sentence. With this knowledge, the models can predict word sequences sequentially, one word at a time.

The key discovery of the LLM was the transformer architecture that was introduced in 2017 (Vaswani and others 2017). The key innovation of the transformer architecture was the introduction of the self-attention feature. This mechanism allows the model to select the key words in the input to pay attention to and deem relevant rather than using the entire input equally (Appendix Figure 2.1).

Transformers are a newer and more powerful type of neural network architecture. They are designed to process sequential data without using recurrent connections. They use an attention mechanism to learn the relationships between parts of the input sequence. This makes them more efficient and easier to train than recurrent neural networks (RNNs). Transformers are better at handling long sequences of data and are also well suited for a variety of NLP tasks. They are easy to scale and easier to train than RNNs.

The computer does not understand words or text; hence, all input words must be converted to vectors before the computer can perform all statistical patterns and mathematical modulization of each step. These are the key steps of a transformer model:

1. **Input embedding:** The input is a sequence of tokens, which are first converted into vectors using an embedding layer, followed by the addition of positional encoding to retain the order of words.

2. **Self-attention mechanism:** The heart of the transformer model is the self-attention mechanism. It allows the model to weigh the relevance of each word in the sequence in producing a representation for each word. In essence, the mechanism captures the context of each word.

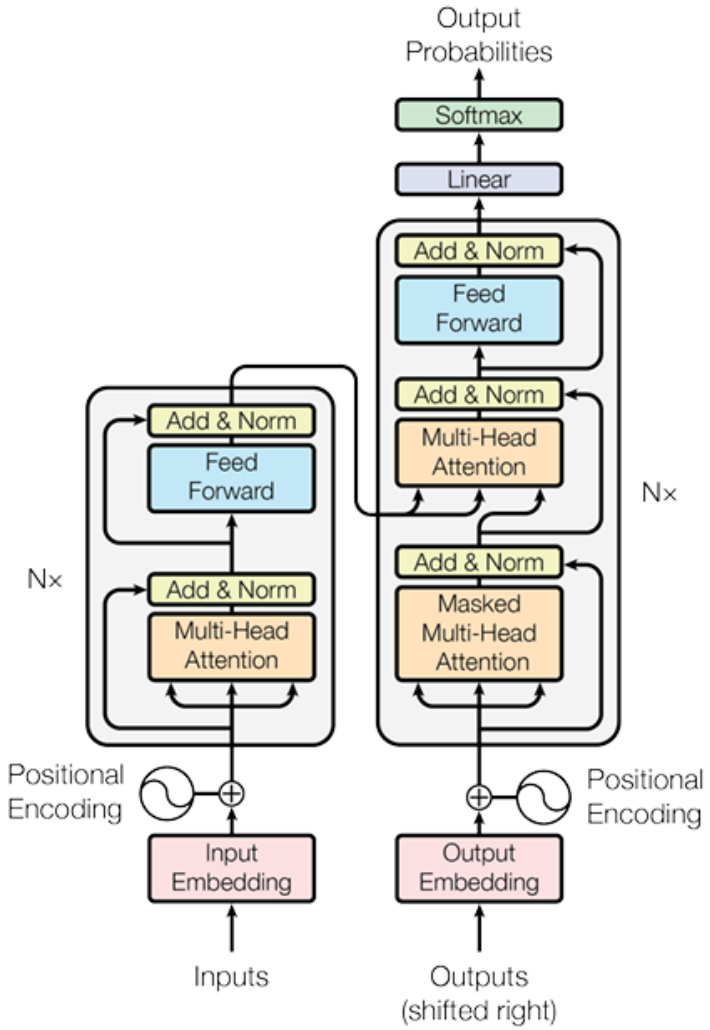
3. **Layer normalization:** After self-attention, layer normalization helps in faster and more stable training.

4. **Feed-forward neural network (FFNN):** Each position in the encoder then goes through a simple FFNN, which transforms the contextualized vectors from the self-attention mechanism.

5. **Stacking of layers:** Steps 2 through 4 are repeated several times. The output of one layer (self-attention + normalization + FFNN) is used as the input for the next one. The number of times these layers are stacked is a parameter of the model and can be changed depending on the complexity of the task at hand.

6. **Output layer:** Finally, the output from the last transformer layer goes through a final linear layer and a softmax activation function for tasks such as language modeling or classification. For language modeling, the output would be a probability distribution over the vocabulary, indicating the likelihood of each word being the next word in the sequence.

Appendix Figure 2.1. The Transformer Model's Architecture



Source: Vaswani and others 2017.

While extremely powerful and capable of creating compelling content, GenAI models do have several limitations:

1. **Understanding context:** While generative AI models like GPT-3 can create grammatically correct and contextually relevant responses, they don't truly "understand" the text in the way

humans do. They are essentially pattern-matching algorithms that have learned to predict, based on the data they were trained on, what comes next in a sequence of text.

2. **Lack of common sense:** Because generative AI models learn from data, they don't possess innate human knowledge or common sense unless it was present in those training data. For example, AI models might not inherently understand that an elephant cannot fit inside a car, unless they've seen similar information in the data they were trained on.

3. **Dependence on training data:** The quality and scope of the training data greatly affect the performance of generative AI models. If the training data are biased, the model's output will likely also be biased. Similarly, if the training data lack certain information, the model won't be able to generate that information accurately.

4. **Control and safety:** It can be challenging to control the output of generative models. They might create content that is inappropriate, offensive, or misleading. This is a significant area of ongoing research in AI safety.

5. **Resource intensive:** Training generative AI models typically requires a lot of computational resources and data, making it inaccessible for individual researchers or small organizations.

6. **Inability to verify facts:** Generative models like GPT-3 don't have the ability to access real-time or current information and can't verify the truth of the information they generate; they can only draw on the knowledge that was available up until the point they were last trained. Applications on top of the models are being developed to perform web searches to look up facts.

7. **Hallucination:** The term comes from the idea that the model is "imagining" or "making up" details that were not in the input and do not accurately reflect reality. Hallucination can be a major issue in tasks where factual accuracy is important, such as news generation or question answering.

Several options exist to help overcome some of the limitations. These options include the following:

1. Integration with knowledge graphs:

A knowledge graph is a powerful tool for storing structured information. It typically represents knowledge in terms of entities (like people, places, objects) and relationships between them. This structured format allows for precise, straightforward queries and can easily link related information. Knowledge graphs are used by search engines (like Google's Knowledge Graph) to enhance search results with semantic-search information gathered from a variety of sources.

Knowledge graphs could be used to complement GenAI models to address some of their limitations. The LLM can be employed to interpret and generate natural language inquiries and responses, while the knowledge graph can be used to deliver factual and consistent information that informs those responses. This combination has the potential to augment the performance of tasks such as these:

- Question answering: The LLM can comprehend the inquiry, while the knowledge graph can look up the accurate response.
- Semantic search: The LLM can interpret the natural language search query and transform it into a structured query for the knowledge graph.
- Information extraction: The LLM can parse unstructured text to extract entities and relationships, while the knowledge graph can store and query this extracted information.

The integration of knowledge graphs with LLMs holds the potential to yield more precise, dependable, and contextually aware responses, marrying the humanlike language generation capabilities of LLMs with the factual consistency and the accuracy of knowledge graphs. Nevertheless, the seamless integration of these two systems presents a complex challenge and remains an active area of investigation within the field of AI.

2. Fine-tuning:

Fine-tuning represents a technique to supplement the training of models like GPT-3, which rely on extensive data sets sourced from diverse origins with more specialized or enterprise-specific data. When the model subsequently generates text, it will produce more focused and accurate output, thereby mitigating the likelihood of spurious or nonmeaningful text. This process facilitates the model's capacity to tailor its output in alignment with the newly incorporated training data.

3. Prompt engineering:

The essence of prompt engineering lies in the deliberate configuration of input structure and content, which is orchestrated to guide and shape the model's output qualitatively. The merits of prompt engineering are manifold—notably, enhancing the precision of the generated text, exercising some degree of control over the output, and, crucially, mitigating inherent bias.

References

Acemoglu, Daron, and Pascual Restrepo. 2019. "Artificial Intelligence, Automation and Work." In *The Economics of Artificial Intelligence,* edited by Ajay K. Agrawal, Joshua Gans, and Avi Goldfarb. Chicago: University of Chicago Press.

ADVERSA. 2023. "Universal LLM Jailbreak: CHATGPT, GPT-4, BARD, BING, ANTHROPIC, and Beyond." Accessed May 26, 2023. https://adversa.ai/blog/universal-llm-jailbreak-chatgpt-gpt-4-bard-bing-anthropic-and-beyond/.

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb. 2018. *Prediction Machines: The Simple Economics of Artificial Intelligence*. Boston: Harvard Business Review Press.

Atreides, Kyrtin. 2023. *Automated Bias and Indoctrination at Scale… Is All You Need*. ResearchGate. http://dx.doi.org/10.13140/RG.2.2.16741.88803.

Boukherouaa, El Bachir, Ghiath Shabsigh, Khaled AlAjmi, Jose Deodoro, Aquiles Farias, Ebru Iskender, Alin T. Mirestean, and Rangachary Ravikumar. 2021. "Powering the Digital Economy: Opportunities and Risks of Artificial Intelligence in Finance." IMF Departmental Paper 2021/024, International Monetary Fund, Washington, DC.

Friedman, Batya, and Helen Nissenbaum. 1996. "Bias in Computer Systems." *ACM Transactions on Information Systems* 14 (3): 330–47.

Dziri, Nouha, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022. "On the Origin of Hallucinations in Conversational Models: Is It the Datasets or the Models?" Paper presented at the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, WA, 5271–85.

Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. "Survey of Hallucination in Natural Language Generation." *ACM Computing Surveys* 55 (12): 1–38.

Nicoletti, Leonardo, and Dina Bass. 2023. "Humans Are Biased: Generative AI Is Even Worse." Bloomberg Technology + Equality. Accessed June 23, 2023. https://www.bloomberg.com/graphics/2023-generative-ai-bias/.

Papenbrock, Jochen, and Alexandra Ebert. 2022. "Best Practices: Explainable AI Powered by Synthetic Data." *NVIDIA Technical Blog*. May 20, 2023. https://developer.nvidia.com/blog/best-practices-explainable-ai-powered-by-synthetic-data/.

Parikh, Ankur, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. “ToTTo: A controlled table-to-text generation dataset.” In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 1173–86.

Retail Banker International. 2023. “Goldman Sachs Experimenting with Generative AI.” March 23, 2023. https://www.retailbankerinternational.com/news/goldman-sachs-experimenting-generative-ai/.

Ullah, Ihsan, Andre Rios, Vaibhav Gala, and Susan McKeever. 2020. "Explaining Deep Learning Models for Structured Data Using Layer-Wise Relevance Propagation." https://doi.org/10.48550/arXiv.2011.13429.

US Consumer Financial Protection Bureau. 2023. “Chatbots in Consumer Finance.” June 6, 2023. https://www.consumerfinance.gov/data-research/research-reports/chatbots-in-consumer-finance/chatbots-in-consumer-finance/.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jacob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. “Attention Is All You Need.” Paper presented at the 31st Conference on Neural Information Processing Systems (NIPS2017), Long Beach, CA, December 4-9, 2017. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

PUBLICATIONS

Generative Artificial Intelligence in Finance: Risk Considerations

NOTE/2023/006