The Alan Turing Institute

The Impact of Large Language Models in Finance: Towards Trustworthy Adoption

This work is part of the FAIR Programme (Framework for Responsible Adoption of Artificial Intelligence in the Financial Services Industry), an EPSRC-funded Prosperity Partnership EP/V056883/1 at The Alan Turing Institute.

The opinions expressed in this publication are those of the contributing authors.
They do not purport to reflect the opinions or views of their organisations or its members.

Corresponding Authors:

Carsten Maple (cmaple@turing.ac.uk)
Alpay Sabuncuoglu (asabuncuoglu@turing.ac.uk)

Project Management and Contact:

Tony Zemaitis (tzemaitis@turing.ac.uk)
Turing - Finance & Economics Programme (FinanceandEconomicsProgramme@turing.ac.uk)

Contributing Authors (Alphabetically Ordered):

Andrew Elliott, The Alan Turing Institute
Andrew Walters, Bank of England
Anna Kharchenkova, Accenture
Fern Watson, Financial Conduct Authority
Gesine Reinert, The Alan Turing Institute
Henrike Mueller, Financial Conduct Authority
Isaac Bowers-Barnard, Accenture
Jagdish Hariharan, University of Warwick
Lukasz Szpruch, The Alan Turing Institute
Marcus Turner, Allen & Overy
Marie Briere, Amundi
Matt Shelley, Accenture
Nataliya Tkachenko, Lloyds Banking Group
Oxana Samko, HSBC
Paul Lickman, HSBC
Pavle Avramovic, Financial Conduct Authority
Praveen Selvaraj, The Alan Turing Institute
Ray Eitel-Porter, Accenture
Sapan Dogra, Accenture
Todd Bose, Standard Chartered
Vijay Jairaj, Standard Chartered

Contents

1 Introduction

It is a pivotal moment for the adoption of artificial intelligence (AI) as large language models (LLMs) introduce ground-breaking advancement in the field. Forecast to become a more than 40 billion USD (£31.5) market by the end of the decade, their ability to process textual data and generate coherent text output has proven highly effective across a diverse range of applications including, but not limited to, healthcare, law, and education. The potential impact of this technology on the finance domain remains relatively unexplored.

Large language models have evolved with the capability of processing vast amounts of linguistic data to generate human-like language responses to prompts or queries, appearing to digest a question, for example, as they predict an expected response by analysing the data.

The Alan Turing Institute and colleagues from HSBC, with support from Accenture and the UK's Financial Conduct Authority (FCA) undertook a study to both explore and build a common understanding around potential opportunities and challenges in the use of LLMs for financial services. This study drew on facilitated discussion between researchers and practitioners from the contributing organisations to determine the key issues for consideration in a survey of existing literature and through a face to face workshop bringing together 43 contributors, representing major high street banks, regulators, investment banks, insurers, consultancies, payment service providers, as well as government and legal professionals working in the sector. This report presents the collective insights of the group to offer a comprehensive view of current development.

Analysis reveals that the financial services sector is living up to its reputation as early adopters of transformative technologies. The majority of workshop participants have begun to employ LLMs to support varied internal processes, and actively assess their potential for market-facing activity in the delivery of advisory and trading services.

Significantly, the majority of participants anticipate LLMs to be integrated into external money economics-related services such as investment banking and venture capital strategy development within two years. They also acknowledged reputational risks and potential for adversarial impacts that must be proactively managed, including for example risks to data integrity driving the accuracy of LLM-generated responses; the opportunity to prompt detrimental or fictional content, or the elevation of new vulnerabilities for sensitive data that can compromise individual privacy. Mitigating privacy risks alongside the integration of effective human-in-the-loop-AI collaboration emerged as the high priorities for enabling more widespread and safe integration of LLMs. Further, while LLMs and their foundational models pose a substantially different opportunity and risk landscape to previous machine learning (ML) techniques, the finance industry's existing agile and robust risk assessment and management tools may offer foundations for facilitating controlled integration.

Key considerations discussed included:

- A requirement to navigate complexities and global considerations that can drive unfair concentration of services in large organisations with the data to support their development, alongside competitive advantage for countries with a favourable regulatory landscape.
- Growing risk of automation bias, and transfer errors between operational functions that could undermine human assistance as reliance on LLMs develop.
- The emergence of security vulnerabilities that come with the complexities of siloed departmental integration of new technologies, including LLMs.
- The importance of articulating and developing human-machine collaboration skills such as prompt engineering and chain of thought and a call for a clearer definition of the relationship between guidelines and human review in decision-making processes supported by LLMs.
- Developments in open-weight and open-source models, and their accessibility for industry and other specific actors, including smaller companies that are providing a platform for collaboration across varied opportunities and challenges.

These insights have led to the tailored recommendations documented in this report for researchers, practitioners, regulators and policymakers to take advantage of this pivotal moment and support the formulation of robust strategies for safe, trustworthy adoption of LLMs. Initially requiring the availability of vast amounts of general data, LLMs now have the opportunity to evolve with more granular task and specific use-case understanding and purpose-specific models. This creates the opportunity to conduct cross-sector analyses of use-case-dependent development that can inform such strategies.

2 Purpose

Working with the support of HSBC, Accenture and the UK Financial Conduct Authority, The Alan Turing Institute conducted a study to understand the benefits and potential impact of LLMs across the financial services sector.

Purpose

The work undertaken reviewed a brief history of the LLM landscape, delved into opportunities and challenges of integrating LLMs in financial services, collated perspectives from experts and explored open-ended questions for achieving safe utilisation of LLMs in financial services.

This report combines the results of an extensive literature survey on the impact of LLMs in banking conducted by The Alan Turing Institute and HSBC in the final quarter of 2023 and insights shared by 43 participants who attended a workshop to build a collective understanding of key considerations, while also examining questions about the likelihood, significance, and timing of the impact of LLMs and related technologies on the financial services sector. It builds on existing bodies of work that, while offering a comprehensive view of the potential implications of developments in machine learning (ML) and AI, may not seamlessly translate to considerations for LLMs and the rapid advancements in their multi-task capabilities.

This research was part of the FAIR Prosperity Partnership established to unlock the transformational benefits of responsible adoption of AI across financial services. The work examined the challenges of using LLMs responsibly, through the lens of the five pillars of the FAIR programme: Robustness and Resilience; Privacy and Security; Fairness and Transparency; Verification and Accountability; and Integration Environment.

3 LLM Development in Financial Services

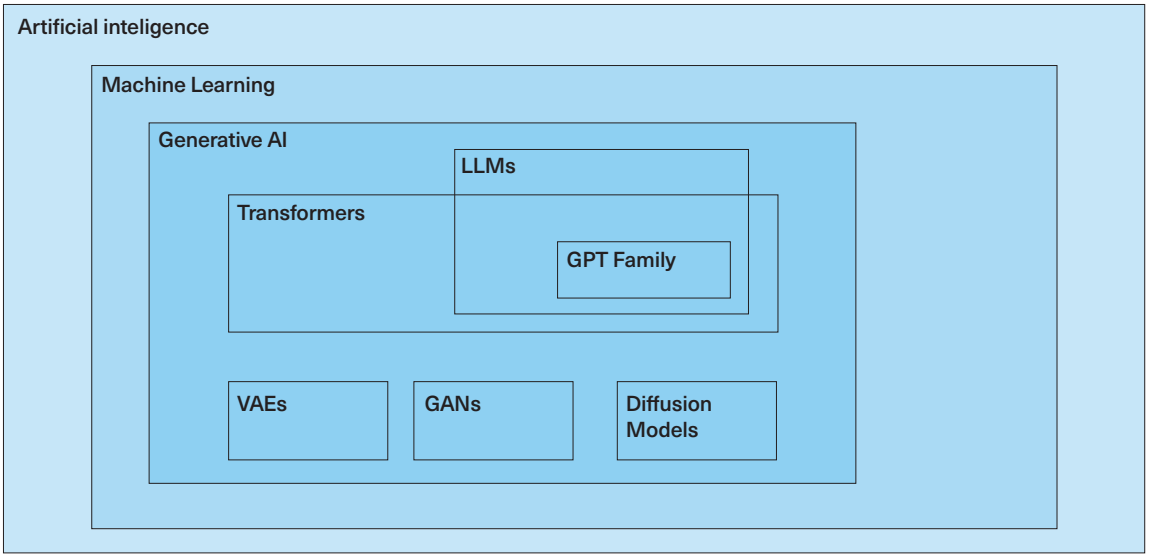
LLMs are the subject of significant interest from governments, regulators and many industry sectors, [6, 7] and are heavily featured in both the academic literature and the popular press. This shared interest underpins a thriving market valued at 10.5 billion USD in 2022 and is anticipated to reach 40.8 billion USD by 2029, demonstrating a compound annual growth rate of 21.4% between 2023 and 2029 [1].

Existing financial services are highly regulated and highly data-driven with machine learning (ML) playing a significant role in many services. A joint Bank of England (BoE) and FCA [8] survey of financial services firms in the United Kingdom (UK) revealed that 72% of respondents use ML applications in their day-to-day applications. The 2023 report of the Financial Policy Committee meeting states that, with the recent advances in LLMs, several financial firms have generated interest in the possibilities of integrating this technology into their services. Some financial companies and service providers to the financial sector have publicly expressed their experimentation with LLMs. However, the report reveals that the current exploration of use cases primarily involves low-risk activities, such as information search and retrieval or generating internal outputs, rather than automating business decisions [9].

LLM Development in Financial Services

A further 2023 survey from UK Finance, a trade association for the UK banking and financial services sector, revealed that more than 70% of participating financial institutions are in the proof of concept (PoC) stage for generative AI solutions, of which LLMs. are an example. One significant investment made by financial software and media company Bloomberg, BloombergGPT [10], has produced a 50 billion parameter model that can be utilised for an array of financial tasks such as news analysis and question answering. With such development, understanding the implications of large-language models in financial services advances the opportunity to set best practices for a critical economic sector, and provide examples that may be relevant to other industries [11].

AI models are generally categorised into two types: predictive and generative. Predictive models are trained to make discrete decisions based on the classification of labelled data. Models with the capability of generating data such as text, images or any kind of data are considered to be generative AI (GenAI) [12]. The most recent examples of Generative AI include image generators (e.g. Midjourney), chatbots (e.g. ChatGPT), code generators (e.g. CodeX), and audio generators (e.g. VALL-E). These systems use complex algorithms and statistical models, applying techniques such as diffusion models, variational autoencoders (VAEs) and generative adversarial networks (GANs) to produce new content mirroring their training data [13]. The foundational models that are often associated with LLMs mainly use transformer architecture to process a vast amount of data.



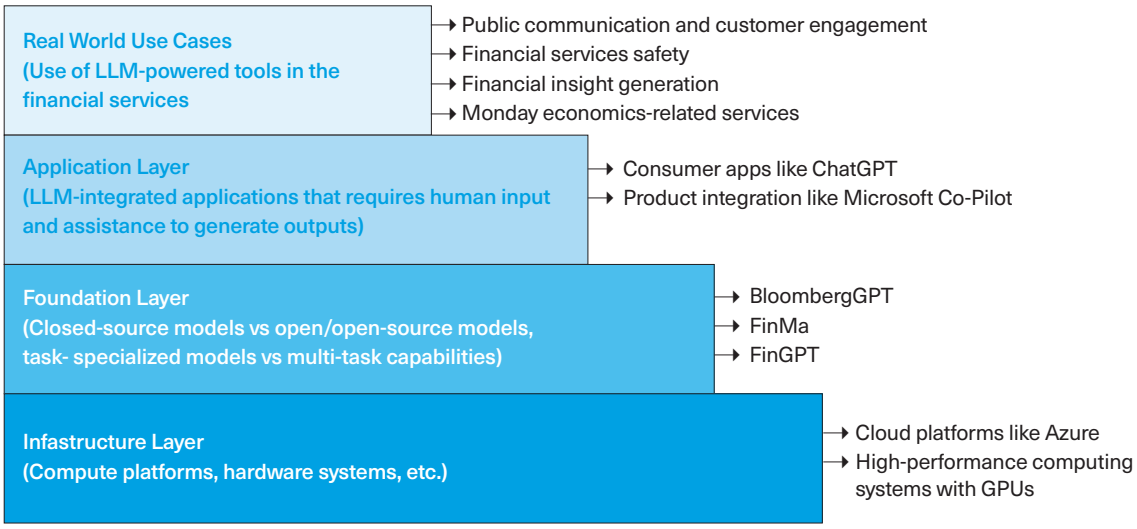
A simplified schematic to illustrate the relation between LLMs, popular GenAI architectures, and broader fields (ML and AI).

A language model is an artificial intelligence system designed to digest text inputs and generate text outputs [14]. In general, language models involve modelling the probability of word sequences to predict the likelihood of future word sequences. The capability of processing and utilising vast amounts of existing linguistic data differentiates large language models (LLMs) from the traditional language models [13].

Large language models are a subset of generative AI, trained on vast amounts of textual data to generate human-like language responses. The development of LLMs is often seen as driven more by increased data and computational power and advances in algorithmic innovation in the model architecture.

3. LLM Development in Financial Services

The landscape of LLMs underwent a significant transformation following the introduction of the transformer architecture by Google researchers in 2017 [15], underpinning models which were initially popularised in the field of Natural Language Processing (NLP), and evolved to be known as LLMs when scaled up to hundreds of millions of parameters. These large models excelled not only in regular benchmarks but also displayed an ability to perform tasks from a single or a few prompts [16]. (See Annex for more detail on the development of the LLM landscape).



A simple depiction of established LLM stacks consists of four layers. Turing researchers have classified real-world use cases into four categories, detailed in Section 4. These financial service categories operate on application layers, with notable examples like ChatGPT and co-Pilot. The foundational layers consist of language models like GPT4, Llama2, Gemini, and Claude. Presently, prevalent open financial LLMs are primarily trained and fine-tuned based on these existing models. The advancements in this domain are made feasible through extensive hardware systems and substantial data management capabilities.

A review of the literature outlines how financial institutions could in theory use LLMs for better decision making such as credit risk assessment, loan approvals and investments. Algorithmic trading is another application that can leverage LLM models to identify potential opportunities in the trading market by using its prediction and analysis capabilities [17]. Providing tailored research and supporting 'next best action' decisions is considered a high-value opportunity to use LLMs in the finance sector. This could, for example, underpin advice on an individual's financial position and investments for a customer's particular circumstances, enhancing value for the customer and deepening the customer relationship for the financial firm. [18].

Significant work is also emerging to advance understanding and opportunities for managing the risks that could evolve with these developments. The financial institution could be liable for poor advice or discrimination, or failing to treat customers fairly, for instance, in a case where there is inherent bias in the outputs. In addition, the vast amount of unsupervised data processed by LLMs brings a risk of privacy [19].

LLM Development in Financial Services

LLMs being developed for the financial sector

BloombergGPT: BloombergGPT is a large language model that was developed by Bloomberg specifically for the financial sector. It is used to generate financial news and analysis, as well as to develop new financial products and services [10].

FinGPT: FinGPT is an open large language model that is trained on a massive dataset of financial data. It can be used for a variety of financial tasks, such as generating financial reports, performing financial analysis, and developing new financial products and services [20].

TradingGPT: TradingGPT is a multi-agent system powered by a LLM with sophisticated layered memory capabilities emulating human traders' cognitive behaviours. This system is designed specifically for the nuances of stock and fund trading markets, enabling it to discern and leverage crucial information from complex layers of financial data to drive informed trading strategies [21].

FinBERT: FinBERT is a pre-trained NLP model to analyse sentiment of financial text. It is built by further training the BERT language model in the finance domain, using a large financial corpus and thereby fine-tuning it for financial sentiment classification [22].

InvestLM: InvestLM, tuned on lLaMA-65B. InvestLM shows capabilities such as processing financial text and provides helpful responses to investment related questions. Financial experts, including hedge fund managers and research analysts, rate InvestLM's response as comparable to those of state-of-the-art commercial models (GPT-3.5, GPT-4 and Claude-2) [23].

PIXIU: Also known as FinMA, provides a comprehensive framework including the first financial LLM based on fine-tuning Llama with instruction data, the first instruction data with 136K data samples to support the fine-tuning and a holistic evaluation benchmark with four financial NLP tasks and one financial prediction task [24].

FLANG: A specialised financial language model, trained on specific financial keywords and objectives, to achieve NLP tasks such as managers' sentiment analysis and financial news classification [25].

BBT-Fin: Chinese financial pre-training language model based on the T5 model. Trained on a large scale financial corpus with approximately 300GB of raw text from four different sources [26].

XuanYuan2.0: The largest Chinese chat model designed for Chinese language in the field of Chinese Finance, built upon the BLOOM-176B architecture and hybrid-tuning approach to mitigate catastrophic forgetting. [27].

DISC-FinLLM: Multiple Experts Fine-tuning Framework to build a Financial LLM. Improves general LLM multi-turn question answering abilities, domain text processing capabilities, mathematical computation skills, and retrieval-enhanced generation capabilities [28].

4 Functional Opportunities

Drawing on the literature survey, researchers categorised key applications of LLMs in financial services under four service categories: (1) Public communication and customer engagement, (2) Financial service safety, (3) Financial insight generation, and (4) Money economics-related services.

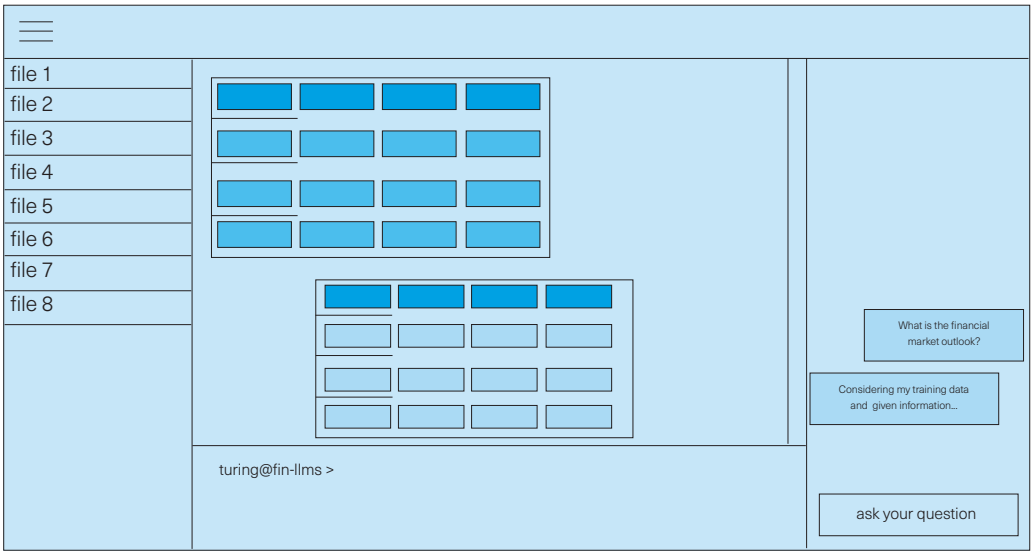
Functional Opportunities

Public Communication and Customer Engagement	Financial Service Safety	Financial Insight Generation	Money Economics-related Services
• Financial Communication • Customer service	• Detecting and preventing fraud • Product development • Risk assessment	• Market surveillance • Market insights and reports • Business finance data insights • Personal investment insights • Generation of aggregate reports	• Investment banking • Treasury optimisation • Private equity and venture capital strategy development • Asset allocation

Public Communication and Customer Engagement include services such as financial communication (simplifying technical jargon for public communication), and customer service. In marketing and customer service, chatbots have been used to improve customer experience and transaction or sales conversion rates [29]. These chatbots are capable of handling customer inquiries, providing support, and personalising interactions more effectively than traditional customer service approaches. LLMs can process and transform the sector-specific jargon used in financial markets to enable better interpretation of market language [30]. Further, they can lead to the development of financial literacy education materials for customers and the wider general public.

Financial Service Safety encompasses various services, including fraud detection and prevention, market and trade surveillance, and risk assessment of financial products. Leveraging their capacity to process extensive transactional data, LLMs can help advance identifying patterns and anomalies that may signify fraudulent activities or financial crimes [31]. Additionally, integrating LLMs into compliance platforms enables compliance officers to swiftly pinpoint potential issues, for example, insider trading within vast communication data, thereby enhancing the efficiency of surveillance operations [32]. Moreover, human-in-the-loop approaches can assist developers in adhering to security and privacy best practices throughout the code-writing process [33]. This integrated approach ensures a comprehensive and cohesive strategy for safeguarding financial services.

Financial Insight Generation includes services such as market surveillance, market insights and reports, business finance data insights, personal investment insights, and generation of aggregate reports [34]. The use of LLMs improves the efficiency in compiling and analysing financial data to produce comprehensive reports on market trends, and company performance, and can offer intricate financial analysis for risk assessment and portfolio management. LLMs can extract business insights from a wide range of structured and unstructured data, providing financial institutions with more unique decision-making foundations. This can include personalised responses with robo-advisors using LLM capabilities to identify user intentions more accurately compared to previously used models, and thereby enhance investment advice and personalised experience. [20, 33].



An illustrative interface of a potential LLM-powered financial application. LLMs can process vast amounts of data and generate outputs in a structured way to increase productivity in financial insight generation. In this illustrative example, we see a chatbot integrated interface that automatically processes the given data to generate meaningful outputs.

Money economics-related services includes services such as investment banking, treasury optimisation, private equity and venture capital strategy development, and asset allocation [35]. LLMs can significantly streamline the due diligence process for mergers and acquisitions and Initial public share offerings (IPOs). The ability of LLMs to analyse and interpret complex financial documents can potentially enhance the accuracy of valuations and identifiy potential synergies, informing strategic advisory services. For treasury operations, LLMs can be instrumental in cash management and risk assessment. The predictive analytics provided by LLMs could help treasurers forecast cash flows more accurately and devise effective hedging strategies, thereby improving the financial resilience of the organisation [36]. In capital markets, LLM-powered services can inform trading decisions and asset allocation strategies to enhance traders and investors ability to align their strategies with market conditions and their risk profiles.

5 Workshop Insights: Functional Considerations

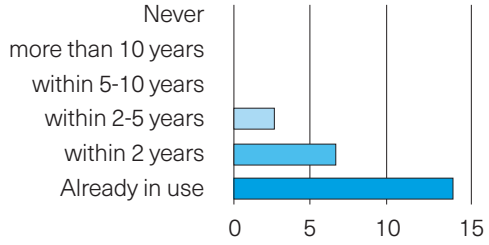
A workshop structured to capture individual viewpoints and personal experiences from 43 contributors, drew on experience from high street banks, regulators, investment banks, insurers, consultancies, payment service providers, government, and legal professionals. Their collective insights covered considerations specific to functional service areas, including the pace of integration, potential opportunities and risks; and responses to open-ended questions designed to clarify ambiguous concepts and develop granularity.

Individual Utilisation of LLMs in Work-related Tasks

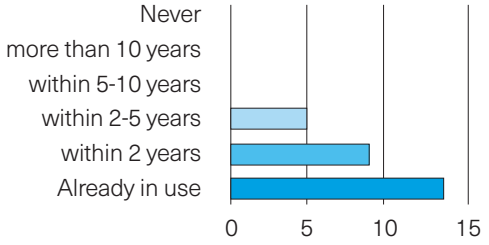
Questions related to participants' personal utilisation of LLMs in work-related tasks elevated understanding of personal preferences and utilisation of the LLM systems within individuals' working environment . The findings indicate cautious adoption of LLMs among participants with high utilisation when they can build active and interpretable collaboration with LLMs. Just over half of participants, 52%, leverage these models to enhance performance in information-oriented work tasks, while 29% employ them to boost critical thinking skills. Additionally, 16% utilise language models to break down complex tasks, and 10% leverage these tools to improve team collaboration. A significant minority, 35% said that they do not currently incorporate any form of language model into their tasks. Participants who use LLMs stated that they use them for mainly risk-free tasks with heavy human assistance, including text summarisation, literature overviews, increasing the speed of analysis, and reinforcing decision-making processes by identifying grey areas.

Functional Perceptions of LLMs

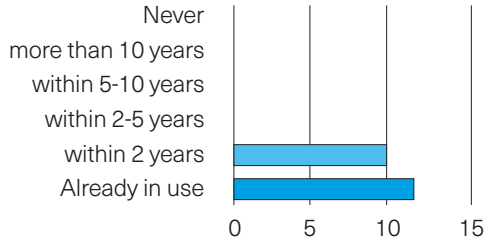
The earliest integration of LLMs in public communication and engagement services



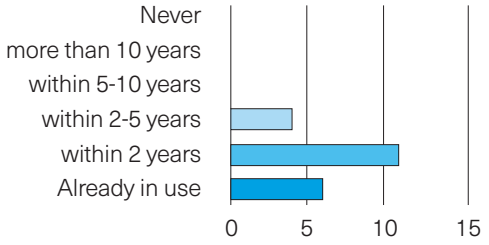
The earliest integration of LLMs in financial insight generation services



The earliest integration of LLMs in financial service safety



The earliest integration of LLMs in "money economics"-related services



[1]In this part, the main research questions were aligned with the recently published report, Microsoft New Future of Work, to interpret the characteristics of participants from the finance sector, with the findings revealed in the report. The four categories (Improving the performance in information work tasks, boosting critical thinking, breaking down complex tasks, and improving team collaboration.) are obtained from this report. More details about our research methodology can be found in Annex Section 1 and at the link: https://doi.org/10.5281/zenodo.10804392.

5. Workshop Insights: Functional Considerations

When asked to anticipate the pace of integration, participants were in agreement with the expectation that the integration of LLMs and related systems would occur in all functional service areas well within five years. Most areas had already experienced some level of use: Money-economics related services were the only category not already being advanced by the majority of the respondents. Their forecasts by functional service area are as follows:

Integration of LLMs in public communication and engagement services: The majority of respondents stated that financial institutions have already integrated LLMs and are utilising them in a variety of services internally. For the most part, LLMs were being used within training scenarios to improve customer facing skills, rather than for direct public contact. Some institutions are still in development or testing phases and expected to integrate these systems within two years.

Integration of LLMs in financial insight generation services: This area demonstrated a similar trend to that identified for public communication and engagement services. The majority of respondents stated that financial institutions have already integrated LLMs and are utilising them in a variety of services internally. However, more respondents expect a slower adoption of LLMs in these services compared to public communication and engagement services.

Integration of LLMs in financial service safety: As with the previous application areas, the majority of respondents state that financial institutions have already integrated LLMs and are utilising them in a variety of services internally. Given that this service area has limited front-facing applications on the customer side, it is expected to be integrated faster with a level of human assistance internally. Integration of LLMs in money economics-related services: The money-economics related services are the only category that was not being integrated by the majority of the respondents. Despite this, most respondents expect it to be integrated within two years.

Integration of LLMs in money economics-related services: The money-economics related services are the only category that was not being integrated by the majority of the respondents. Despite this, most respondents expect it to be integrated within two years.

Broader Opportunities

In parallel to the function-specific developments discussed, the following themes emerged multiple times offering more granularity into how promising application areas of LLMs in financial services are developing:

Operational: The finance sector is already establishing systems to enhance productivity through rapid analysis of large amounts of text or co-pilot-like tools to process semi-structured or unstructured information. These operational tasks include streamlining decision-making processes, risk profiling, benefit quantification, and prioritisation, improving investment research, and back-office operations.

Enhancing human-machine interaction: As part of an AI system that analyses data, LLMs can be integrated to enhance the natural language interface flows. This can include a wide array of applications from dictation to embedded assistants. Such integration would simplify internal processes and reduce the complexity of internal knowledge -intensive tasks such as the review of regulations to suggest governance controls and compliance tools. They could in theory also accelerate credit analysis, client due diligence, and transaction monitoring.

Financial advisory: Personalised robo-advice fed by a broad range of data, and enhancing financial literacy are the two most promising opportunities advanced by workshop participants. Integration of LLMs could enhance strategic and advisory services and help experienced professionals do more. Leveraging advanced NLP capabilities, financial institutions could also integrate multiple media types, such as images, into comprehensive internal assessments. A team of experts could leverage LLMs to synthesise publicly available information to inform asset allocation decisions or develop actionable insights by processing signals from a broad range of input data to be interpreted by experts.

Financial literacy: The generative capability of LLMs could be used to power financial literacy educational environments and improve financial inclusion with personalised support based on an individual's literacy levels. Children could learn about investments and pension, for example, or actions could be taken by directing customers to the right channels.

6 Risks: LLMs in Financial Services

Risks posed by LLMs are a significant limiting factor for financial institutions. They manifest at different levels, including data, model, and governance, and have the potential to be amplified within interconnected systems.

Risks: LLMs in Financial Services

Establishing deep trust with customers is paramount as clients expect their investments to remain secure and confidential. Examples of risks posing a threat to the financial institution's brand include exposure of sensitive data; data, system or organisational manipulation by malicious actors; and the expression of views by LLMs that are not in alignment with corporate policy. The literature survey was used to identify potential LLM risks in three categories that informed their examination by the workshop participants :

(1) Risks arising from the need for vast amounts of unsupervised data,
(2) Risks associated with the complexity of the models, and
(3) Risks linked to social behaviours and human values.

Data-related risks	Model complexity-related risks	Social behaviours and Human Values
• Bias Privacy • Data transparency and security • Violation of intellectual property	• Lack of explainability • Reasoning errors • Susceptibility to various attacks	• Alignment • Information hallucination • Toxic linguistic • Environmental impacts • Open vs close source impacts

Data related risks

Bias: Ensuring accurate information is crucial in finance. Financial institutions could be liable for poor advice, discrimination or failing to treat customers fairly, in cases where there is inherent unfair bias in the LLM outputs. Rigorous risk and testing assessments are needed to prevent toxicity and bias [37]. They can unintentionally demonstrate areas of bias present in their training data, leading to discriminatory or inaccurate outputs [38]. LLMs may exhibit social biases and toxicity [39, 40] during the generation process, resulting in the production of biased outputs.

Privacy: The mix of public and private data in finance necessitates careful scrutiny of data sources to ensure privacy. Training LLMs on sensitive data raises privacy issues, as models can inadvertently memorise and reveal private information or provide accurate statistical information about private data. Techniques such as federated learning and the use of LoRA (Low-Rank Adaptation of Large Language Models is a training technique to reduce the number of trainable parameters) weights, which support training and fine-tuning of models should be explored to enhance privacy protection [41, 42].

Data Transparency and Security: Different strategies for mitigating risks associated with LLM release, range from open sharing to restrictive application programming interface (API) access or no access, highlighting the need for the development of robust security controls and compliance frameworks [27, 43-46].

Intellectual Property (IP) Risks: Reliance on extensive datasets, which might include copyrighted material, inherently increases the risk of IP infringement. The use of copyrighted works, or material otherwise subject to intellectual property rights in training data poses significant risks of intellectual property infringement, both at the point of training and at the point of use. There is a potential for lawsuits and ethical dilemmas when copyrighted or licensed content by is reproduced by LLMs [47, 48].

6. Risks: LLMs in Financial Services

Model complexity-related risks

LLMs Explainability: The complexity of the models with billions of parameters, make it difficult to understand and interpret their decision-making mechanisms [49-51]. Lack of transparency hinders insight into how specific inputs lead to outputs. [52, 53]. This poses a particular challenge in the context of regulatory transparency obligations which arise in a number of different contexts, including AI-specific regulation, data protection laws, and financial services regulation.

Susceptibility to Various Attacks: LLMs are vulnerable to adversarial attacks such as misleading prompt injections, 'jailbreak' attacks, and data poisoning. LLMs are sensitive to prompts, especially adversarial prompts [54], which alter evaluations and algorithms and impact their robustness.

Reasoning Errors and Quality of Outputs: LLMs have limited abilities in abstract reasoning [55] and are prone to confusion or errors in complex contexts [56]. LLMs can make mistakes in logical reasoning due to ambiguities in prompts or limitations in understanding complex operations. Further, fidelity and quality of LLMs can change going forward as increasingly their inputs are the output of other LLMs.

Struggles in Specific Tasks: LLMs can exhibit restricted proficiency in discerning semantic similarity between events [57] and demonstrate substandard performance in evaluating fundamental phrases [58]. LLMs exhibit subpar performance and encounter challenges in accurately representing human disagreements [59]. LLMs have limitations in incorporating real-time or dynamic information [60], making them less suitable for tasks that require up-to-date knowledge or rapid adaptation to changing contexts. That being said, specific techniques, such as retrieval augmented generation (or RAG) are improving an LLM-based system's capabilities in this regard.

Social behaviours and human values

Alignment: Understanding, interpreting, and acting in accordance with human values is a major goal in LLM development. However, existing LLMs generate socially undesirable outputs with hallucinations, jailbreaking and data leakages. As a result, LLMs can demonstrate misaligned behaviours which can cause reputational, allocational, or societal level harms[61].

Information Hallucination: LLMs can generate factually incorrect or fictional but believable information, a complex issue related to the training process, dataset, and architectural design [62]. This is one of the most significant limitations to widespread deployment of LLMs by financial institutions, particularly for customer services. LLMs may manifest credibility deficits [63], potentially giving rise to fabricated information or erroneous facts within dialogues [54, 64].

Toxic Linguistic: In linguistic contexts featuring non-Latin scripts and limited resources, LLMs manifest suboptimal performance [36, 65, 66]. Furthermore, generative LLMs consistently display proficiency levels below the expected standards across various tasks and languages [65]. LLMs demonstrate susceptibility when processing visual modal information [67]. Furthermore, they have the capacity to assimilate, disseminate, and potentially magnify detrimental content found within the acquired training datasets, frequently encompassing toxic linguistic elements, including offensive, hostile, and derogatory language [39].

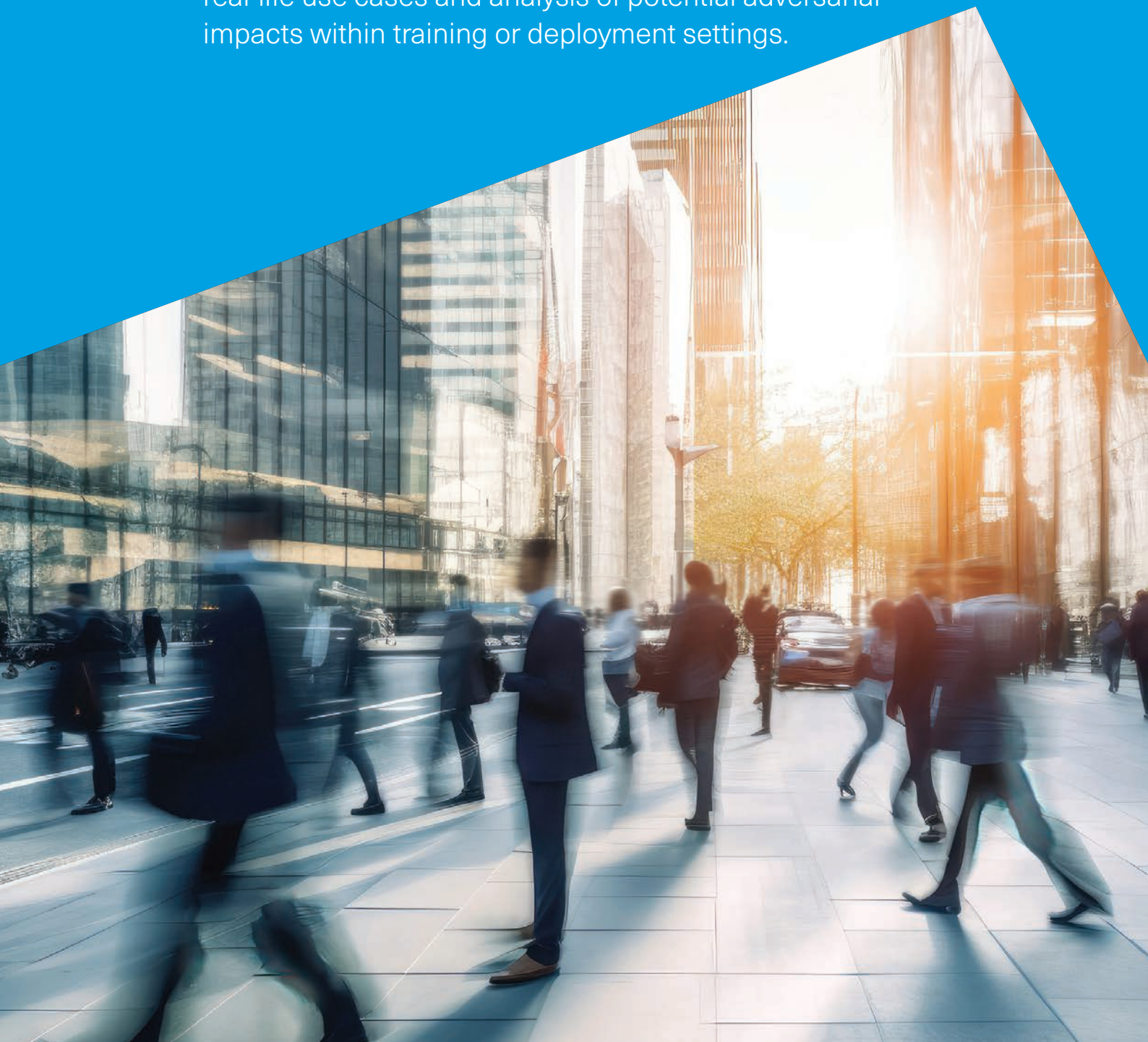
Environmental: Due to high computational power requirements, training LLMs demands high energy and water consumption [68]. This high consumption of water is primarily due to the cooling process of data centres, which necessitates a massive amount of water to regulate the servers' optimal temperature. Apart from water usage, the training of LLMs demands a considerable amount of electricity. The training of OpenAI's GPT-3 alone was reported to have resulted in the release of 502 metric tons of carbon, hundreds of years of energy for an average American household.

Competition among for-profit organisations: OpenAI was founded around the premises of having a large-scale AI company operating as a not-for profit organisation [69]. However, with the evolution of scale and nature of investment required, it converted to a for-profit organisation AI [70]. Open-source oriented organisations like HuggingFace are supporting the growth of AI with a more participatory approach. Being a for-profit organisation can introduce the risk of excessive control by the investors, and of decoupling development from issues arising from the general public [71].

Openness and Open-Source Approach: Although the emphasis on open data, open models, open-source code, and open education can foster collaborative problem-solving and innovation, it can also reveal backdoors for white-box attacks [72], requiring assessment of potential trade-offs in the open-source implementations.

7 Workshop Insights: Risks

With the Integration of LLMs introducing new risks and opportunity for adversarial impacts, workshop participants were asked to rank their perceived levels of risk by each functional service area. Public communication and customer engagement, and financial service safety functional areas, both areas identified by the majority as actively integrating LLMs, were considered to present the greatest levels of risk. This suggests that their assessment of risk was based on some real-life use cases and analysis of potential adversarial impacts within training or deployment settings.



Workshop Insights: Risks

In the context of LLM integration, researchers defined harms as unintended consequences and potential adversarial impacts covering reputational, legal, and societal consequences. In all services, nearly all respondents selected legal and reputational harm as the most likely and most impactful harm in the integration of LLMs in financial services. This presents expectation for more investment in risk management governance controls, including inserting fire breaks between the LLM output and use of that output in the business, to ensure appropriate review and sense checking of outputs and prevent reputational and legal harm before wide-scale deployment in these areas.

Participants' concerns centred around two main areas of potential loss: Loss of confidence in financial services should a catastrophic failure occur, or heavy reliance on LLMs create a systemic issue; and the loss of core skills if financial institutions reduce their workforce with the use of LLMs. Further, using this disruptive technology in a work environment that processes sensitive and privileged data prompted recognition of more consideration than deployment in other areas. A lot of LLMs are being used for meeting notes, call notes, interview notes and the like. These can contain privileged information which can be difficult to track, particularly if within the control of a departmental silo or transferred across areas.

Privacy risk appeared as one of the highest priorities in the integration of LLMs. Nearly half of the participants were concerned about the privacy vulnerabilities introduced with LLM systems and potential data leakage. Such challenges involve legal considerations and compliance with various regulatory frameworks. A major challenge is determining which jurisdiction and, accordingly, which regulation applies if the developer and deployer of a LLM, or the data subjects and their data used by the LLM are in different countries. Further, retiring an LLM raised concerns about preserving customer history files. Strategies for retention and transfer need to be established to ensure seamless transitions and compliance with data protection regulations.

The potential for inaccuracy within the generated text (hallucination and bias) was the second concern but with human assistance in the decision-making process, participants did not consider this a major risk. However, the level of integration raised concerns about the risk of automation bias; whereby significant reliance on LLMs could potentially introduce an adverse impact on human judgement and control. Further, the participants acknowledged that the integration of new models may disrupt organisational workflows, leading to errors passed between operational silos. Measures for establishing trust in model outputs is essential to prevent such adverse impacts and ensure accuracy and reliability in decision-making as reliance on automated processes develop.

Other areas of risk highlighted include:

Concentration risk: In the current scale of LLMs, only a handful of providers can develop, run and maintain these models. Concentration risk and opportunity for data asymmetry may arise if there is insufficient competition among providers of LLMs, or the large organisations with access to data to develop them internally. The high cost of training LLMs in terms of data gathering and training stages also resulted in domination in the LLM space by big tech companies. This can be exacerbated when providers are concentrated in a single jurisdiction or benefit from a favourable regulatory landscape.

7. Workshop Insights: Risks

Lack of assurance and traceability: Information about the extent of data used for training is currently almost impossible to retrieve or validate. This traceability becomes a bigger concern when LLMs are developed in operational silos. Further, support of a complex business logic with multiple LLM agents working together can create hidden feedback loops that can cause systematic risks. Participants highlighted the need for comprehensive testing methodology and statistical assurances, that are continuous, and part of the production process for launching these models.

Unknowns and complexity: There are multiple unknowns related to LLMs specifically their impact on security and privacy compliance. Considerations for protecting reputation are significant, with great potential for public hypersensitivity to AI mishaps, and difficulty in proving something is not the case when a firm is accused of AI-related harms, as exampled by the Apple Card discrimination accusation[84]. Further, GDPR becomes a factor when personal data is embedded, as would be the case with hyper-personalised services, raising questions such as how to extract this data in response to a subject access or deletion request.

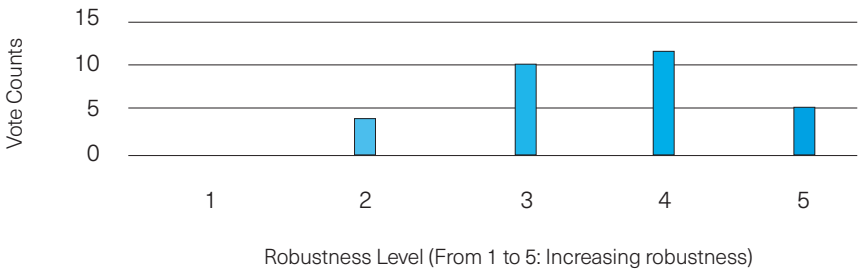
Further Considerations

Robustness of internal guidelines: Participants were asked to rate their own internal guidelines for LLMs. Only 19% of the respondents said their internal guidelines on the use of LLMs were highly adequate. On average, they rated the robustness of their internal measures at 3.7 on a scale of 1 to 5. While most institutions have established internal requirements, policies and procedures governing the use of LLMs, participants highlighted the need for heightened governance considerations related to privacy, data security, and the accuracy of responses. Specifically, there is a call for a clearer definition of the relationship between these measures and human review in decision-making processes empowered by LLMs. Further, given their multi-tasking capabilities, it was recognised that a company exhibiting a high level of governance maturity in predictive ML usage, for example, can still expect to encounter a myriad of challenges when implementing LLMs.

The current regulatory focus for financial services is on rules and principles that deliver the right outcomes for consumers, markets, and the economy. They are therefore designed to be highly flexible and adaptive, rather than prescriptive about how outcomes are achieved. As a highly regulated industry, the use of LLMs in daily tasks was expected to be regulated in a similar fashion, with a focus on outcomes. Financial institutions, in comparison to other sectors, introduce an added layer of sensitivity regarding privacy to mitigate the risks associated with misinformation. Their internal policies are predominantly tailored to regulations relevant to specific territories and markets, aligning with ethical, legal, and customer duty implications.

A crucial requirement for supporting internal guidelines is the provision of clearer examples illustrating the maturity level of LLM-based systems and their limitations. This necessity extends to offering guidance on human factors and providing examples of user journeys, applicable to both internal and third-party development. The guidelines should include warnings regarding the limitations inherent in LLMs and delineate when and why utilising LLMs would be preferable compared to other similar tools, such as keyword search or other available functions in office applications, that may serve the purpose.

Robustness of internal guidelines for utilising LLMs in their own organisation



Some participants also noted that clarification for the use of R&D (experimentation) and real-life use cases is necessary as they highlighted a lack of internal benchmarks for transition of ownership and proportional guidelines between the development and product phases of LLM integration.

Geopolitics and regulatory restrictions: Many representatives shared concerns about potential copyright breaches. With varying guidelines and jurisdictional differences across countries, existing regulations lack harmonisation. This creates a notable area of risk and uncertainty. Environmental costs, including carbon costs and the role of regulation in different geographies also raised concern around competitive challenges, particularly between the United States and Europe. Concerns about concentration risk highlighted that different levels of access to information are concentrated in some jurisdictions.

Lacking AI and financial literacy in the users: There is a widening gap in AI literacy between model creators and organisational leadership. This gap may lead to misinformed investment decisions, as leaders may struggle to evaluate AI technologies effectively, hinder strategic alignment, and prevent the successful integration of AI into business operations. This highlighted a need for executives to ensure they are sufficiently skilled and understand the risk of not understanding risks.

The same gap occurs between the developers and the general public. AI literacy with some level of financial knowledge can prevent herding behaviour by enabling individuals to understand risks, think critically, and maintain a long-term perspective. Awareness of historical and behavioural biases and informed decision-making based on market conditions can help individuals to resist speculative trends.

8 Towards Safe Adoption

With the increasing speed of AI development, achieving safety and trustworthiness has become a priority for both public and private institutions. The National Institute of Standards and Technology (NIST) has defined trustworthy AI systems through seven key characteristics: validity and reliability, safety, security and resilience, accountability, and transparency, explainability and interpretability, privacy enhancement, and the fair management of harmful bias. Each of these characteristics involves numerous subcomponents, contributing to the complexity of achieving a shared understanding and realisation of truly trustworthy AI.

Workshop participants discussed these critical factors, including whether they had clear definitions, to explore what good would look like in financial services. Touching on both the use and development of LLMs, discussions revealed existing foundations—guidance on the robustness of existing predictive models, and banking ethics councils, for example—that may be useful in their advancement of LLMs. They also delved into approaches for tackling complexities such as concentration risks, potential bias in unstructured data, and compliance with privacy regulations, acknowledging that the utility of LLMs inherently increases the collection and organisation of data. Further, examined considerations and defining requirements for explainability, suggested value in developing granularity, with explainability defined at various levels.

The level of granularity appeared as a pivotal concept essential for ensuring the safe integration of LLMs across diverse discussion topics. Lack of granularity in both training data and model capabilities is the primary challenge for managing both security and privacy risks, and for defining robustness checklists.

A growing focus on the analysis of human behaviour was also imperative, as underscored in both security and privacy discussions. Understanding human behaviour to inform and implement proactive guidelines was advocated as a strategic consideration for achieving a commendable level of security. Crucially, this should advance recognition of consumer vulnerability, with models equipped with indicators to identify potential vulnerabilities, and reflect good awareness of malicious activity such as phishing attacks where criminal actors disguise themselves as banks. Another area covers analysis of how human interaction will develop with LLMs. This is particularly crucial as demonstrated by the effectiveness of ELIZA, one of the earliest natural language processing computer programmes (developed from 1964 to 1967) at engaging and thereby compelling individuals to reveal sensitive information.

Robustness and Resilience

Participants emphasised that robustness is use-case dependent, and that in some areas it is easier to evaluate what good looks like. Robustness relates to generalisability in many ways, and in some scenarios, they can be used interchangeably. ISO/IEC TS 5723:2022 [85] defines robustness as the "ability of a system to maintain its level of performance under a variety of circumstances." Resilience depends on building robust processes and systems so that they can withstand adversity or recover fast. Participants emphasised the need for a test environment that can map external and internal threats and develop the outputs iteratively throughout a development process.

Most financial services are bound to strict robustness checklists due to either regulations such as the Gramm-Leach-Bliley Act [75], or to prevent reputational and legal harms. These methods might include verification techniques, inspecting noise factors and failure modes. However, defining the measures of performance when the tasks are not atomic (isolated), was elevated as a challenge due to the multi-task capabilities of LLMs. The disruptiveness of these models comes from their ability to comprehend a wide variety of data and generate outputs in a variety of styles, while the robustness evaluation requires running atomic use-case evaluations to prove the system demonstrates a certain level of performance in different conditions for different sets of inputs.

The finance sector has already developed guidance on the robustness of predictive models. The discussion explored the opportunity of developing techniques to build atomic levels of LLM capabilities, so that they can be evaluated following the existing robustness and resilience checklists. Overall, the group suggested identifying the main differences between LLMs and existing ML models could accelerate the creation of the right checklist elements.

Reviewing the existing robustness terminology coming from the research and comparing it with the terminology surrounding advancing understanding of current risks in the LLMs could also help practitioners define measurable characteristics. Aligning with that of existing regulations and internal guidelines was considered critical to achieving robust test designs. For example, a generative AI-specific term, hallucination, referring to factually incorrect or fictional outputs, presents one of the main challenges in terms of achieving robust development pipelines. However, hallucination as a term is hard to grasp and not directly transferable to test cases.

8. Towards Safe Adoption

The participants focused on measuring reproducibility and repeatability, sensitivity testing, and explainability to initiate the evaluation of robustness. For the latter, they posited that transparency of evaluation and assurance techniques is more important, than transparency in terms of explainability of the model architecture, underlining the need to achieve a level of explainability that allows for auditing. For internal processes, it was accepted that the risks related to imperfect model outputs can be mitigated through human monitoring. Defining levels of robustness and linking them with the possible autonomy levels depending on the use case could therefore be an alternative approach to defining what good looks like in the robustness definition. Checking validity evaluates good. Ideally, this should draw on many individuals' summaries, and cover testing of a breadth of systems followed by evaluation techniques using humans such as red teaming.

Data Asymmetry

Data asymmetry between big tech and financial services firms emerged as a growing concern. The accessibility of data that underpins LLMs has potential to underpin competitive advantage with visible impact on consumer quality. FCA recently had an open call on this issue [76]. With bigger firms revealing they have already started to test their systems with their existing data sources, concern was raised about the representativeness of smaller firms' data for fine-tuning LLMs for their use cases. Discussions covered parameters for sharing data, whether this should be bilateral between firms or regulatory driven to be system-wide, and the implications for the duty of care for banks that prevents easy sharing of data. This included concern that customer trust and reputation could be adversely affected should regulators require open sharing of anonymised data. By contrast, the group explored the potential for expanding digital sandboxes that already exist to allow for such sharing; and the development of synthetic datasets with open availability to support different use cases.

Attention was given to the defining incentives for big financial institutions: Bigger firms may have a significant competitive advantage, with a lot to give and less to receive. While it was suggested that regulations could force their hand, organic incentives may also accelerate and support firms to naturally collaborate. Different types of companies are innovating and advancing varied exciting ideas using data. Big banks may be able to learn from the smaller banks, for example on their data governance practices, and innovative ways of working. Smaller companies can also contribute enriched data and innovative data management and maintenance solutions.

The group suggested that government-collected datasets could also be made available publicly to mitigate data asymmetry. Legislative amendments, like The Data Protection and Digital Information Bill (the DPDI), and provisions on data sharing could expand the scope of data accessibility, drawing parallels with the principles of open banking. This reflects a need to address geographical differences as nations develop advantage with regulatory policy around data privacy, copyright, and other measures. The European Commission as a regulator, for example is steering access to data toward smaller firms.

The Case of Open Banking

In the recently released Staff Working Paper no 1059, they investigated the early evidence from open banking data to understand customer data access and fintech entry.

Open Banking (OB) refers to the growing practice of empowering bank clients to share their financial transaction details from their bank accounts with various financial service providers. As an illustration, OB allows a bank client to utilize a mobile application to effortlessly disclose their bank account history to a prospective lender (who can evaluate their income and spending patterns for credit underwriting) or to a financial management app (facilitating effective money management).

The following excerpt from a staff working paper published in February 2024 (Appendix E) reveals some observed benefits of open banking:

"In general, implementing OB policies results in heightened fintech participation across various financial products. UK consumers utilize OB for financial advice and credit products, with these applications linked to increased financial knowledge and credit accessibility, respectively. Within the UK SME sector, companies influenced by OB are more inclined to establish fresh lending connections, particularly with non-banking entities. This tendency is primarily observed in SMEs with pre-existing lending ties, contradicting the financial inclusion objectives of OB policies but aligning with the distributional forecasts from our model."

* Staff Working Papers describe research in progress by the author(s) and are published to elicit comments and to further debate. Any views expressed are solely those of the author(s) and so cannot be taken to represent those of the Bank of England or to state Bank of England policy. This paper should therefore not be reported as representing the views of the Bank of England or members of the Monetary Policy Committee, Financial Policy Committee or Prudential Regulation Committee.

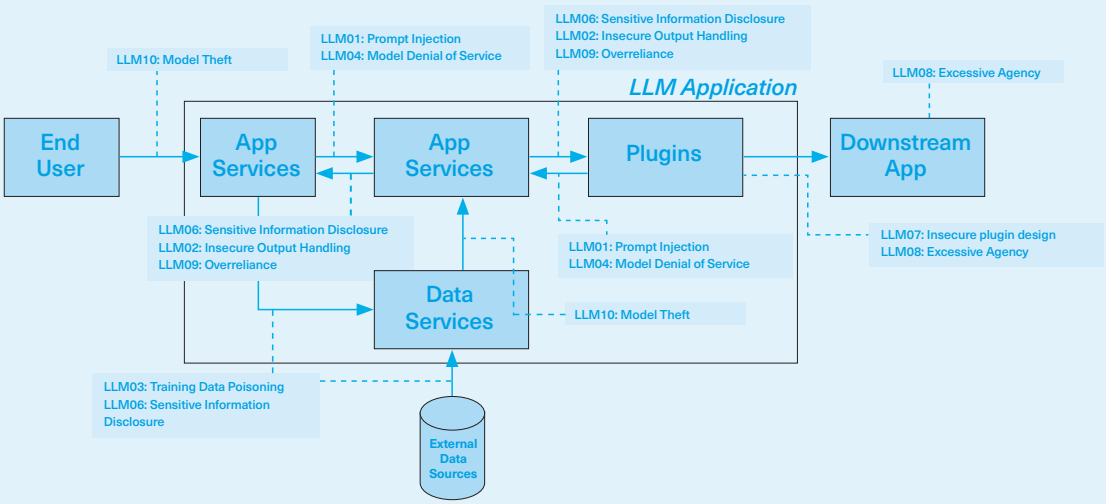
https://www.bankofengland.co.uk/working-paper/2024/customer-data-access-and-fintech-entry-early-evidence-from-open-banking

Security

A recent report from OWASP [77] highlights the security challenges and attack surfaces by the integration of LLMs in business applications. The key security concern in this report is prompt injection attacks, which become a wide and complex issue considering enterprise-level complex business logic. Where and how LLMs ingest input data, highlight a fundamental need for a secure process to prevent sensitive data from being compromised. Complex business logic suggests an attack surface open to multiple attacks including prompt injection, insecure plugin design, and remote code execution.

8. Towards Safe Adoption

OWASP Top 10 Attack Strategies for LLMs



The Open Worldwide Application Security Project (OWASP) recently published the Top 10 attack strategies in a typical LLM utilisation pipeline.

1. Prompt Injection: Crafted inputs can lead security breaches.
2. Insecure Output Handling: Unvalidated outputs can invite security exploits.
3. Training Data Poisoning: Tampered training data can impair models generate inaccurate, or ethically problematic responses.
4. Model Denial of Service: Overloading LLMs can cause service disruptions.
5. Supply Chain Vulnerabilities: Compromised components, services or datasets can breach the integrity.
6. Sensitive Information Disclosure: Disclosure of sensitive outputs can result in legal consequences or a loss of competitive advantage.
7. Insecure Plugin Design: Untrusted inputs and lacking access control risk severe exploits.
8. Excessive Agency: Unchecked autonomy can lead to jeopardizing reliability, privacy, and trust.
9. Overreliance: Critically assessing LLM outputs is key to mitigate security vulnerabilities, and legal liabilities.
10. Model Theft: Unauthorized risks theft, competitive advantage, and dissemination of sensitive information.

*Source: https://owasp.org/www-project-top-10-for-large-language-model-applications/

Understanding this context workshop participants highlighted three priority areas for managing security risks:

Third-party vendors: Even in cases where an LLM is developed to mitigate security risks, the third-party infrastructure which enables use could introduce security risks. The hosting arrangements for LLMs and the evaluation of the security stance of third-party vendors (for example through an assurance framework or underwriting) are therefore of concern. While a third-party assurance framework may present robust due diligence requirement for vendors, their limitations should be acknowledged, particularly in scenarios where smaller organisations are in negotiation with big, multinational providers and have limited influence. Further security risks arise from the architecture in which LLMs are executed on top of these third-party systems.

Open vs closed source models: Software developers often face a significant challenge when choosing external software for their systems considering the security of third-party systems. One way of increasing scrutiny on the software supply chain is introducing "software artifacts" with "trusted metadata" to ensure software supply chain integrity and enhance security in the software development lifecycle. Model signing in this sense can be applied to achieve trust in the open model integration process. Currently, the "officially signed" models are mostly released as closed models (sharing weights with a license agreement, or providing API endpoints), which makes closed source models more secure. However, the legal implications of using less secure, open models, built on top of other open architectures, present a topic of ongoing debate. As a security-first industry, the models are expected to be officially signed to provide a level of assurance. In this regard, assurance techniques and supply chain security techniques are gaining importance.

The trade-off between security and openness: Some security considerations can result in the over-protectiveness of data which can restrict innovative applications. This is influencing a security vs. 'good' trade off where models that are proprietary and developed in-house for security reasons, may be less effective than more open models. The Open Banking initiative proved itself illustrating that while data can be shared among the stakeholders with the consent of the consumer, it can be also advance a secure environment. This depends on the design of the access control, however, there are currently many unknowns in the LLMs to manage. Institutions are therefore focusing on inbound and outbound security aspects of LLM data and saying that inbound is easier to protect, which maintains the current focus on internal use.

Thinking outside the customer data, it was also noted that using LLMs in the product development and coding process may introduce vulnerabilities for proprietary information. Should a company lose control of its privileged data or should information which is a trade secret be disclosed to a third party via an LLM, this may have an impact on preserving intellectual property rights, including patentability.

8. Towards Safe Adoption

Privacy

In the legal space, there is active debate on the potential privacy impacts of LLMs in different domains, including financial services. Although GDPR and DPDI cover some of the fundamental concerns, privacy is both a legal and ethical concern, with the consumer perspective on the collection and use of their data the subject of ongoing discussion.

Our group indicated that ethics councils from banks may have a say in this, and that by default, LLM use may currently go to these councils. To manage this, they suggested the scope of these councils and ethics evaluation frameworks may need re-evaluation to facilitate social and technical perspectives. They also emphasised a need to map GDPR requirements against LLM capabilities to inform potential updates that may be required of this legislation. Noting that DPDI helped in open banking and is being expanded, they highlighted a need for measures to address assurances of accuracy, defamation by usage of personal data, and the role of the data controller, alongside the need for institutions to determine risk appetite in these areas.

The group also cited opportunities to apply technical privacy-enhancing techniques such as differential privacy, and the use of sandboxed environments to advance innovation and assessment of varied use cases and business models.

Fairness

Effective fairness monitoring demands both technical and social capabilities in the development teams: Lacking one of these capabilities might result in inadequate evaluation of fairness. In LLM development, researchers and developers actively seek answers to open questions on defining technical definitions and metrics for effective evaluation of fairness in a multidisciplinary approach. In LLM development, most institutions try to understand if measuring fairness for an LLM is different for any other model. and whether it is harder given the natural language output of a model. A key challenge is the inability to do attribution of underlying protected characteristics with LLMs to assess fairness which is not quantitative. They queried whether opportunities could be developed for facilitating comparison of outcomes with current interactions, for example how interacting with a chatbot vs getting a model response from a service powered by LLM impacts the level of fairness in a response.

To accommodate these challenges, participants advocated that fairness and bias-free principles need to be baked into human agent scenarios. Participants acknowledge that some financial data is already biased. For simple tabular data addressing this can be straightforward, but it was unclear how to de-bias unstructured textual data. The group also raised concern that bias in machine systems could be more significant than bias in human decision-making. LLMs should, therefore be subject to a higher standard of scrutiny regarding bias and careful evaluation with human assistance.

Current largely internal use is implemented with heavy human guidance, after an analysis of the use case that determines if the LLM benefits the purpose. They also noted that some areas are not necessarily favourable for LLMs.

Explainability

While the lack of explainability in decision making is preventing industry from using LLMs in many applications, workshop participants advocated that having accurate but intuitively explainable models are more important than having complete explainability. They also noted that proportionality and therefore different levels of explainability and transparency may be appropriate for different use cases.

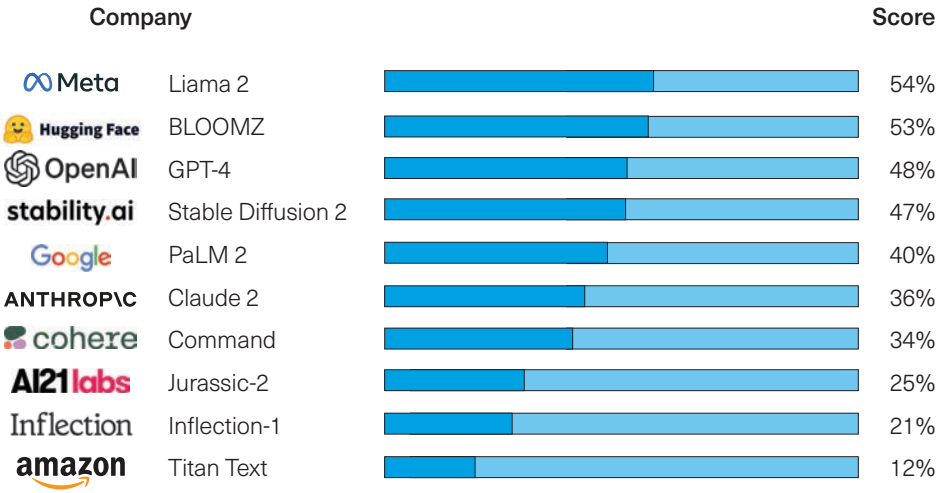
The focus here was on facilitating auditability: The details of training data could be open, or a corporate-level certification or signature could be added to foundational models to make them auditable. This would allow experts, governments and relevant stakeholders to define approaches for unravelling the complexity involved.

With decision-making processes having significant effects for clients, financial institutions do require a higher level of explainability than other sectors. Currently, there is limited regulation on enforcing the explainability of the models in decision-making. Building on long-standing data protection principles, regulators are increasingly focused on transparency obligations in the context of LLMs (see, for example, obligations for providers of general-purpose AI models under the new EU AI Act). Discharging these transparency obligations is likely to pose a particular challenge in the context of LLMs where traditional explainability methods are difficult to achieve. Institutions can utilise auditing and conformity assessment to measure company competency. Further, employee training and attention to detail could be advanced to support institutions' development of an intuitive understanding of how models can behave in different use cases and contexts, and underpin the development of auditability.

8. Towards Safe Adoption

An Initial Effort to Indexing Model Transparency

The Foundation Model Transparency Index developed by Stanford University [78] is a comprehensive assessment of the transparency of foundation model developers.



Foundation Model Transparency Index Total Scores, 2023
Source 2023 Foundation Model Transparency Index

The index follows a detailed approach to calculate indicators that assess transparency across three critical dimensions:

1. Upstream resources such as the utilisation of data, labour, and computing power in building foundation models.
2. Model details like size, capabilities, and potential risks associated with the model itself.
3. The downstream use of the model such as distribution channels, usage policies, and the geographical areas affected.

In this index, major developers, including OpenAI, Google, and Meta, are scored against these indicators concerning their flagship foundation models, such as GPT-4, PaLM 2, and Llama 2. The scoring process involves assigning points based on the extent of information provided by developers across these indicators. This systematic evaluation aims to establish the level of transparency in each dimension, fostering a standardised and comparative analysis of different foundation model developers.

Although this index provides a good starting point to analyse the transparency across different models, academics and practitioners in the area also criticise the impact of individual indicators and model selection criteria. Additionally, in this fast-developing domain, new open-source models such as Olmo (AllenAI) and Gemma (Google) are missing. However, the indicators can help financial institutions to evaluate their risk management frameworks.

Accountability (and Transparency)

Accountability emerged as potentially more important than explainability and transparency. It was linked to the maintenance of customer trust, drawing on understanding of the difference between the recommendations of an advisor and an online chatbot, and a history of success coming from black box decisions. The concept of generating a traceable decision-making trail was mentioned as a way to provide transparency and accountability, allowing for the examination of the decisions made by the models and the mitigation strategies implemented across potential fairness and accountability concerns.

Financial Institutions current Internal use is allowed only after addressing specific questions and meeting internal benchmarks. Institutions opting for third-party services assume accountability to clients, requiring an internal assessment of satisfaction with established benchmarks. While there are established accountability policies and internal standards, discussions explored whether there should be different standards for LLM given the breadth of their use, their risks, and the teams involved. Participants pointed to a need to clearly define the processes and interactions within a legal document that supports oversight. They put emphasis on development of frameworks of accountability based on use-cases, supported by a model owner who is accountable for defined roles, defined tasks, procurement, and ensuring risk assessments are passed.

It was clear that accountability for LLMs have many aspects and involve various teams with everyone working from their understanding of the best standards or practices possible for their area. LLMs are at a level can be seen as software managed by IT, but their multi-purpose nature limits full analysis of accountability concerns in this function. Individual accountability for decision-making in loans for example would be different than system-wide accountability. In the case of GDPR, for example, the data controller has the highest level of accountability. Responsibility will depend on numerous factors, including whether the LLM is used to provide services to consumers, in which case more protective consumer legislation than that afforded to business will come into play.

8. Towards Safe Adoption

Integrity

When discussing integrity, there was a significant overlap with other concepts such as fitness, propriety, consistency, and adherence to "strong moral principles." The issue of integrity may raise questions about the connection between a company's brand value and the uniform application of them. The ability to adhere to established rules is a key requirement, along with measurable criteria like quality persistence, yet applying such well-defined standards of operation to LLMs poses a significant challenge.

Overall, the group advocated that integrity is measured by customer trust, noting that building a trustworthy product is distinct from building overall trust. This distinction holds true across various elements of an integration such as the user interface (UI) interaction with the data and AI model, as well as the personalisation steps. The dynamics vary based on use cases and technological advancements. Different outcomes may emerge and can result in legal and reputational harms as illustrated by incidents such as when the US parcel delivery firm DPD's chatbot swore at a customer.

Determining the ethical use of LLMs raises questions about when it is appropriate to utilise such technology and when specific types of data or anchors can be employed. Ethical considerations align with principles governing the responsible use of AI data. Assessing adherence to these principles serves as a measure to check for integrity in the application of responsible AI practices.

The reliability and ethical integrity of data are intricately linked. While ethical integrity is often discussed, the aspect of data integrity is sometimes overlooked in the context of Large Language Models (LLMs). A disconnect arises between third-party developers and companies regarding the quality and integrity of training data for LLMs, with companies emphasising a greater need for data integrity.

Current preference for "human in the loop" over complete human replacement underscores the importance of maintaining integrity. The group emphasised that it remained crucial to monitor and examine the ways in which humans and machines interact, while oversight of these interactions is needed to uphold the integrity of the collaborative process.

Skills

Fundamental literacy and critical thinking skills with a comprehensive understanding of domain-specific knowledge are still the most important skills and knowledge set for the future workforce of finance. Understanding the policy and regulations and their potential impact on their respective areas is critical in an era where disruptive technologies appear constantly.
With the fast adoption of LLMs into their current services, institutions have started transforming their recruitment and internal training processes. Human-machine collaboration skills have gained importance including prompt engineering and chain of thought in the context of LLMs. In this process, decision-making and considering the probability of error and inaccuracies become much more important.

In the early days of Wikipedia, the average user tended to use the information directly without checking the main source. This changed as users developed the instinct to question what they were consuming. Similar instincts will be needed to develop for LLMs. The number of LLM users is much higher than the number of developers of LLM-based systems leading financial institutions to focus educational training specifically on the effective use of LLMs. However, in some cases, training, governing and maintaining LLMs can be assessed to take considerable amount of time and money compared to realised or anticipated productivity gains.

Education on privacy and security and misinformation within the context of LLM use has been integrated into the cybersecurity and digital skills training courses. In addition, understanding consumer duty in product development processes that involve LLM systems, have been developed. The group, however, identified a gap in training for executives who will need to understand these models to support the development of accountability and assignment of responsibilities.

9 Concluding observations

The recent and rapid rise of large language models (LLMs) has influenced shared excitement across academia, industry and governments. Their unprecedented success in a wide variety of tasks, such as generating realistic conversations, extracting meaningful knowledge, and translating between languages has fuelled this excitement alongside significant public attention.

Concluding Observations

The finance sector shares this excitement, and as illustrated in this study, are already integrating or assessing LLMs into various application areas, from improving customer experience to streamlining market and trade surveillance, generating financial insights, and detecting fraud and suspicious activities. We can anticipate that LLM integration into financial services is likely to continue and develop rapidly: Workshop participants agreed that LLMs could yield significant benefits for departmental functions. They also highlighted opportunity to resolve complexities with multiple LLM agents increasing capability in unstructured data processing, for example, and reasoning skills as agents bring together task-specific knowledge.

Such developments require proactive management of core challenges, particularly privacy challenges that come with a myriad of impacts such as reduced traceability of the knowledge used between operational silos; amplified bias or hallucinations through layered communication, and the use of unfiltered, unstructured data. Delving into the potential evaluation and mitigation techniques thematically, the professionals in the workshop elevated comprehensive opportunity for collaboration. These included sharing known incidents, developing knowledge and best practices to build trust, and the development, maintenance and fine-tuning of foundational or frontier models for specific purpose.

Like any black-box predictive model, LLMs are subject to challenges related to security, privacy, fairness, robustness and lack of explainability. The finance sector has been using ML techniques and faced similar challenges in different application areas. Utilising and building upon the existing risk assessment frameworks could reduce the complexity of the transition.

Further, a detailed examination of factors such as the level of open-sourcing, differing impacts on industry and specific actors, such as smaller companies are ongoing conversations to be advanced. There have been developments on both closed-source models such as BloombergGPT and open models such as FinGPT. As a security-first industry, the models are expected to be officially signed and expected to provide a level of assurance. In this regard, assurance techniques and supply chain security techniques are gaining importance across both approaches.

Overall financial institutions undertake careful planning, weighing both environmental and economic costs for the maintenance, inference, and development of LLMs. These considerations are also dependent on geopolitical positions and existing regulations: Significant variations may occur in access to energy resources, data centres, and other critical infrastructure in different locations. Fundamentally, cutting-edge LLMs are not yet reliable enough to be deployed beyond low-risk, internal-facing financial use cases. They are not yet suited to automate material decision-making. Our participants, however, illustrated the potential for a shift toward the development of specialised, cost-effective, and environmentally friendly LLMs, now often referred to as little or small LLMs as an essential next step. Throughout their discussions, they delved into opportunities for developing achievable granularity, emphasising nuanced levels of precision and detail that can be attained to support such a shift.

10 Recommendations

Though it is a highly regulated industry, the finance sector is known for being an early adopter of cutting-edge technology, and discussions regarding the incorporation of LLMs into financial services show the potential to provide best practices for other sectors. Previously, successful implementation of technology to support open banking and anti-money laundering measures have provided insights into technology deployment risk assessment and management in other sectors.

Recommendations

This work facilitated an initial effort to build collective understanding in achieving safe integration of these systems by bringing together participants from a variety of financial institutions including major high-street banks, regulators, investment banks, insurers, payment services providers, government and legal professionals.

Drawing on the collective insights and reflecting the anticipated pace of integration of LLMs across functional areas, this study points to significant opportunity and imperatives for sector-wide collaboration that can inform robust strategies for the safe adoption of LLMs. Two areas stand out as priority:

1/ The development of use-case dependant, sector-wide analyses of LLM assessments: Collaboration to share and develop advancing levels of granularity would extract knowledge currently developing at the implementation level and largely within functional silos that could inform cohesive approaches for moving forward. In particular, workshop participants encouraged calls for inputs and forums to share safety concerns, adversarial incidents, and best practices that can foster collective learning. Exploration of emerging techniques, such as the use of synthetic data and privacy enhancing technologies would benefit from this cross-sector perspective, while some of the global concerns discussed including competition risks from data asymmetry would require it.

2/ Exploration into opportunities emerging with open-source models: The growing academic interest in open models specialised in financial tasks has the potential to advance LLMs in line with data protection requirements as these models can be utilised in local containers, evaluated internally and maintained with incremental updates. Mitigating security and privacy concerns was identified as one of the highest priorities to enable the widespread integration of LLMs. This requires collaborative research and development efforts from technical, legal, and ethical perspectives that can be supported with open-source projects.

Specifically, the study underlines the following opportunities for the contributing stakeholder communities:

> **Academic Community**: The integration of LLMs into financial services requires careful analysis of the specific use cases, supporting their internal and external use, existing guidelines and risk management scenarios. Structuring the research questions based on specific use cases, while considering current regulations, best practices, and internal analysis strategies, can empower researchers to yield high-impact research outputs. Further, development of well-documented benchmarks can accelerate the systematic evaluation of LLMs across use cases using similar data sources and metrics. Additionally, conducting cross-sector analyses of risks and corresponding implementation experience would contribute to the formulation of robust strategies for the safe adoption of LLMs (or more broadly, AI models with multi-task capabilities) by identifying best practices for risk assessment and mitigation from a multi-disciplinary perspective.

10. Recommendations

Financial Institutions: Financial institutions offer a significant opportunity to take a lead in the development of and maintenance of trustworthy and safety by design principles for LLMs by collating current assessments that are already in motion and reflect a multi-disciplinary perspective. The opportunity covers current efforts to examine ethical, environmental, human-value alignment, privacy and security, model and data risks.

It also draws on potential to utilise current capacities for auditing, conformity assessment, and other measures to support an intuitive understanding of how models can behave in different use cases, and thereby inform traceable decision-making trails, and levels of auditability that can underpin confidence in LLM-supported decisions. Analysis, definition and articulation of the relationship between a firm's guidelines and human review in decision-making processes empowered by LLMs is also recommended to elevate and federate lessons being learned. Further, participation in sector-wide and cross-sector initiative should be developed to bring together perspectives across large and small organisations and advance consensus for industry-level requirements such as fair access to data or the development of principles for governance that map to regulatory and ethical obligations.

Regulators & Policymakers: The value to be extracted from LLMs is wide ranging, requiring ongoing discussion around both the often complex implications for compliance across varied regulations, particularly privacy regulations, and emerging developments in techniques such as the use of synthetic data, that may call for more regulatory scrutiny. Further such wide-ranging potential calls for the development of a coherent multi-lateral set of agreed rules for an internationally harmonised and effective regulatory landscape that can underpin innovation, safety and fair access to opportunity as the use LLMs mature.

This study set out to explore and reveal a comprehensive overview of the opportunities and challenges of LLMs in financial services, putting a focus on the likelihood, significance, and timing of their advancement. The combined findings from the workshop and earlier secondary research not only present such a view, but they also reveal opportunity to develop with collective understanding of the considerations for the development of trustworthy implementations, and begin to set out a path toward their achievement.

11. Annex

Methodology

In the final quarter of 2023, the Turing and HSBC produced a report on the impact of LLMs in banking as a result of an extensive literature survey. This research was part of the FAIR Prosperity Partnership to understand the challenges of using LLMs responsibly, through the lens of the five pillars of the FAIR programme (Robustness and Resilience; Privacy and Security; Fairness and Transparency; Verification and Accountability; and Integration Environment).

This report on LLMs in banking has become the basis for the current study by opening a broader discussion on LLMs in financial services. The team from the Turing and HSBC conducted an extensive literature survey to identify the current applications and risks of LLMs in financial services.

In the next stage, they invited attendees from major high-street banks, regulators, investment banks, insurers, payment services providers, as well as government and legal professionals working in financial services to hold a consensus-building workshop. Forty-three participants attended this workshop. Before the workshop, the literature survey findings were shared with the participants. During the workshop, they were asked questions about the likelihood, significance, and timing of the impact of LLMs and related technologies on the financial services sector and beyond. The research team focused on understanding individual and functional perspectives on the use of LLMs in financial services, expanding the discourse to consider implications across the five pillars of the FAIR program. They also asked questions related to individual and functional perceptions and collected answers using the Slido [79] platform. Notetakers took notes during the open-ended questions and informal conversations.

The team analysed the participants' responses and notetaker notes following a thematic analysis process and aligned the analysis findings with the previously identified opportunities and challenges. As a result, this report aligning with the pillars of the FAIR programme seeks to understand the scope, timing of adoption, barriers, benefits and potential impact of LLMs across the financial services sector. It presents a broad scope of LLMs in financial services by bringing in the viewpoints of the workshop participants.

A Brief History of Language Models

The first language models were developed in the 1950s and 1960s. These models were rule-based and relied on handcrafted linguistic rules and features to process language. One of the oldest instances of an AI language model is the ELIZA, which made its debut in 1966 at MIT [80]. Although these models were limited in their capabilities and their ability to handle complexities in different natural language processing (NLP) tasks [81], a recent pre-print also claimed that these models still can surpass modern LLMs in certain tasks [82]. In the 1980s and 1990s, statistical language models were developed. These models used probabilistic methods to estimate the likelihood of a sequence of words in a given context. They were able to handle larger amounts of data and were more accurate than rule-based models [83]. However, they still had limitations in their ability to understand the semantics and context of language [84].

The next major breakthrough in language modelling came in the mid-2010s with the neural language models [85]. These models used deep learning techniques to learn the patterns and structures from large amounts of text data. The first neural language model was the recurrent neural network language model (RNN-LM), which was developed in 2010. RNN-LM was able to capture the context of words and produce more natural-sounding text than previous models [86].

In the early stages of 2014, state-of-the-art models employed encoder-decoder architectures for translation tasks. In 2015, Google introduced the first large-scale neural language model called the Google Neural Machine Translation (GNMT) system [87]. This model was trained on massive amounts of bilingual text data and was able to achieve state-of-the-art performance on machine translation tasks. However, this approach faced limitations, especially with longer sentences, as the encoder struggled to compress all information into a fixed-length vector. In 2017, Vaswani [15] introduced the transformative transformer architecture. They challenged the notion that the success of previous models lay in bidirectional RNNs, proposing that attention mechanisms played a more crucial role. Their model utilised multi-layer perceptrons (MLPs) and a parallelised attention mechanism, incorporating multiple 'attention heads' for improved results in machine translation. The Transformer was able to learn the longer-term dependencies in language and allowed for parallel training on multiple Graphical Processing Units (GPUs), making it possible to train much larger models. Although the transformer architecture has remained largely consistent over the years, a minor modification suggested by Xiong [88] involved placing layer normalisation before the attention layers, leading to improved convergence without the need for a learning rate warm-up stage originally deemed necessary.

In this brief history of language models, numerous models have emerged. Language models with many parameters and utilising significant processing power are collectively referred to as "Large Language Models" (LLM) [89]. The landscape of LLMs underwent a significant transformation following the introduction of the transformer architecture by Google researchers in 2017 [15]. The transformer architecture, initially popularised in the field of Natural Language Processing (NLP), was dubbed LLMs when scaled up to hundreds of millions of parameters, in models such as BERT. The term LLMs likely emerged to differentiate these models from their smaller predecessors and to highlight the changes resulting from their increased scale. These large models excelled not only in regular benchmarks but also displayed an ability to perform tasks from a single or a few prompts [16].

11. Annex

LLMs in the Finance Sector

For this study report, we define 'financial services' as services provided by financial institutions in the area of finance. Financial institutions are institutions that primarily offer financial products. These institutes include organisations working on accounting, banking, financial planning, insurance, investments and pensions, tax, regulation, financial markets (LSE), and legal services. We do not include organisations whose main offering is not a financial service. For example, Apple has a major offering of Apple Pay, but it is not typically called a financial institution as it is not primarily what it offers [90]. Following this definition, the key groups analysed in this report include retail banks, commercial banks, insurance, investment banks, asset managers and fund managers.

Bank of England and FCA's AI Public and Private Forum paper on AI and ML (February 2022) [91, 92] listed some opportunities and risks of using recent AI techniques on a variety of financial services that can potentially benefit from AI: "from customer services to consumer credit, anti-money laundering and anti-fraud analytics to investment management".

Banks have always innovated fast and embraced the NLP paradigms (e.g. chatbots) early on. However, these models had limited capabilities, following rule-based flows and using Language Models (LM) to intelligently parse and understand the given query [93]. Recently developed LLMs are magnitudes more capable in performance. As a result, these models have the potential to make significant advancements in the finance industry with applications ranging from financial NLP tasks, risk assessment, algorithmic trading, market prediction and financial reporting [94].

LLMs such as BloombergGPT [10], a 50 billion parameter large language model trained on large diversified financial corpus, have revolutionized financial NLP tasks such as news classification, entity recognition and question answering. By utilising the huge amount of financial data available, these models have advanced capability to enhance customer services drastically by efficiently handling customer queries and providing them with excellent financial advisory. In addition, LLMs are being used for risk assessment and management, by analysing past market trends and data, it is able to identify potential risks and provide mitigation steps through different financial algorithms [11].

References

[1] PR Newswire, "Large Language Model (LLM) Market Size to Grow USD 40.8 Billion by 2029 at a CAGR of 21.4% | Valuates Reports," https://finance.yahoo.com/news/large-language-model-llm-market-151500260.html, 2023.

[2] The Royal Society, "Post-graduate science students break Large Language Model guardrails at Royal Society AI safety event," https://royalsociety.org/news/2023/11/ai-safety-red-teaming, 2023.

[3] A. Mislove, "Red-Teaming Large Language Models to Identify Novel AI Risks," https://www.whitehouse.gov/ostp/news-updates/2023/08/29/red-teaming-large-language-models-to-identify-novel-ai-risks/, 2023.

[4] CADE, "International Conference on AI and the Digital Economy," https://warwick.ac.uk/fac/sci/wmg/news-and-events/events/wmgevents/cade2024/, 2024.

[5] The Alan Turing Institute. "Synthetic Data," https://www.turing.ac.uk/research/interest-groups/synthetic-data, 2024.

[6] IDST, "Iconic Bletchley Park to host UK AI Safety Summit in early November," 2023.

[7] AAIP, "Laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts," *Eur Comm*, vol. 106, pp. 1-108, 2021.

[8] Bank of England, and Financial Conduct Authority, "Machine learning in UK financial services," https://www.bankofengland.co.uk/report/2022/machine-learning-in-uk-financial-services, 2022.

[9] Bank of England, "Financial Policy Summary And Record Of The Financial Policy Committee Meeting on 21 November," https://www.bankofengland.co.uk/-/media/boe/files/financial-policy-summary-and-record/2023/fpc-summary-and-record-december-2023.pdf, 2023.

[10] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, and G. Mann, "Bloomberggpt: A large language model for finance," arXiv preprint arXiv:2303.17564, 2023.

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

[12] F. García-Peñalvo, and A. Vázquez-Ingelmo, "What do we mean by GenAI? A systematic mapping of the evolution, trends, and techniques involved in Generative AI," 2023.

[13] M. U. Hadi, R. Qureshi, A. Shah, M. Irfan, A. Zafar, M. Shaikh, N. Akhtar, J. Wu, and S. Mirjalili, "A survey on large language models: Applications, challenges, limitations, and practical usage," *TechRxiv*, 2023.

[14] Y. Bengio, R. Ducharme, and P. Vincent, "A neural probabilistic language model," *Advances in neural information processing systems*, vol. 13, 2000.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[16] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, and D. Metzler, "Emergent abilities of large language models," *arXiv preprint arXiv:2206.07682*, 2022.

References

[17] C. Maple, L. Szpruch, G. Epiphaniou, K. Staykova, S. Singh, W. Penwarden, Y. Wen, Z. Wang, J. Hariharan, and P. Avramovic, "The ai revolution: opportunities and challenges for the finance sector," *arXiv preprint arXiv:2308.16538*, 2023.

[18] E. B. Boukherouaa, M. G. Shabsigh, K. AlAjmi, J. Deodoro, A. Farias, E. S. Iskender, M. A. T. Mirestean, and R. Ravikumar, *Powering the Digital Economy: Opportunities and Risks of Artificial Intelligence in Finance*: International Monetary Fund, 2021.

[19] Y. Yao, J. Duan, K. Xu, Y. Cai, E. Sun, and Y. Zhang, "A survey on large language model (llm) security and privacy: The good, the bad, and the ugly," *arXiv preprint arXiv:2312.02003*, vol. 1, 2023.

[20] H. Yang, X.-Y. Liu, and C. D. Wang, "FinGPT: Open-Source Financial Large Language Models," *arXiv preprint arXiv:2306.06031*, 2023.

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

[22] D. Araci, "Finbert: Financial sentiment analysis with pre-trained language models," *arXiv preprint arXiv:1908.10063*, 2019.

[23] Y. Yang, Y. Tang, and K. Y. Tam, "InvestLM: A Large Language Model for Investment using Financial Domain Instruction Tuning," *arXiv preprint arXiv:2309.13064*, 2023.

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

[25] R. S. Shah, K. Chawla, D. Eidnani, A. Shah, W. Du, S. Chava, N. Raman, C. Smiley, J. Chen, and D. Yang, "When flue meets flang: Benchmarks and large pre-trained language model for financial domain," *arXiv preprint arXiv:2211.00083*, 2022.

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

[27] X. Zhang, and Q. Yang, "Xuanyuan 2.0: A large chinese financial chat model with hundreds of billions parameters." pp. 4435-4439.

[28] W. Chen, Q. Wang, Z. Long, X. Zhang, Z. Lu, B. Li, S. Wang, J. Xu, X. Bai, and X. Huang, "DISC-FinLLM: A Chinese Financial Large Language Model based on Multiple Experts Fine-tuning," *arXiv preprint arXiv:2310.15205*, 2023.

[29] M. Company, "The next frontier of customer engagement: AI-enabled customer service," https://www.mckinsey.com/capabilities/operations/our-insights/the-next-frontier-of-customer-engagement-ai-enabled-customer-service, 2023.

[30] M. R. Kabir, and F. Lin, "An LLM-Powered Adaptive Practicing System," 2023.

[31] B. Luo, Z. Zhang, Q. Wang, A. Ke, S. Lu, and B. He, "AI-powered Fraud Detection in Decentralized Finance: A Project Life Cycle Perspective," *arXiv preprint arXiv:2308.15992*, 2023.

[32] E. X. Li, Z. Tu, and D. Zhou, "The Promise and Peril of Generative AI: Evidence from ChatGPT as Sell-Side Analysts," *Available at SSRN 4480947*, 2023.

References

[33] AtomCapital, "Challenges and Opportunities of Applying LLMs in Finance industry," https://atomcapital.xyz/f/challenges-and-opportunities-of-applying-llms-in-finance-industry, 2023.

[34] K.-C. Yang, and F. Menczer, "Large language models can rate news outlet credibility," *arXiv preprint arXiv:2304.00228*, 2023.

[35] FinanceGPT, "Tools Guide," https://financegpt.uk/use-guide, 2023.

[36] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, and W. Chung, "A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity," *arXiv preprint arXiv:2302.04023*, 2023.

[37] T. Tseng, A. Stent, and D. Maida, "Best practices for managing data annotation projects," *arXiv preprint arXiv:2009.11654*, 2020.

[38] P. Schramowski, C. Turan, N. Andersen, C. A. Rothkopf, and K. Kersting, "Large pre-trained language models contain human-like biases of what is right and wrong to do," *Nature Machine Intelligence*, vol. 4, no. 3, pp. 258-268, 2022.

[39] S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith, "Realtoxicityprompts: Evaluating neural toxic degeneration in language models," *arXiv preprint arXiv:2009.11462*, 2020.

[40] A. Parrish, A. Chen, N. Nangia, V. Padmakumar, J. Phang, J. Thompson, P. M. Htut, and S. R. Bowman, "BBQ: A hand-built bias benchmark for question answering," *arXiv preprint arXiv:2110.08193*, 2021.

[41] D. Byrd, and A. Polychroniadou, "Differentially private secure multi-party computation for federated learning in financial applications." pp. 1-9.

[42] Y. Liu, T. Fan, T. Chen, Q. Xu, and Q. Yang, "Fate: An industrial grade platform for collaborative learning with data protection," *The Journal of Machine Learning Research*, vol. 22, no. 1, pp. 10320-10325, 2021.

[43] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, and M. Gallé, "Bloom: A 176b-parameter open-access multilingual language model," *arXiv preprint arXiv:2211.05100*, 2022.

[44] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, and F. Azhar, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[45] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, and A. Askell, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877-1901, 2020.

[46] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, and S. Gehrmann, "Palm: Scaling language modeling with pathways," *arXiv preprint arXiv:2204.02311*, 2022.

[47] C. Novelli, F. Casolari, P. Hacker, G. Spedicato, and L. Floridi, "Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity," *arXiv preprint arXiv:2401.07348*, 2024.

[48] K. Foss-Solbrekk, "Three routes to protecting AI systems and their algorithms under IP law: The good, the bad and the ugly," *Journal of Intellectual Property Law & Practice*, vol. 16, no. 3, pp. 247-258, 2021.

References

[49] D. V. Hada, and S. K. Shevade, "Rexplug: Explainable recommendation using plug-and-play language model." pp. 81-91.

[50] Y. Gao, T. Sheng, Y. Xiang, Y. Xiong, H. Wang, and J. Zhang, "Chat-rec: Towards interactive and explainable llms-augmented recommender system," *arXiv preprint arXiv:2303.14524*, 2023.

[51] A. Uchendu, "REVERSE TURING TEST IN THE AGE OF DEEPFAKE TEXTS," The Pennsylvania State University, 2023.

[52] Q. V. Liao, and J. W. Vaughan, "AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap," *arXiv preprint arXiv:2306.01941*, 2023.

[53] G. Vilone, and L. Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence," *Information Fusion*, vol. 76, pp. 89-106, 2021.

[54] K. Zhu, J. Wang, J. Zhou, Z. Wang, H. Chen, Y. Wang, L. Yang, W. Ye, N. Z. Gong, and Y. Zhang, "PromptBench: Towards Evaluating the Robustness of Large Language Models on Adversarial Prompts," *arXiv preprint arXiv:2306.04528*, 2023.

[55] G. Gendron, Q. Bao, M. Witbrock, and G. Dobbie, "Large Language Models Are Not Abstract Reasoners," *arXiv preprint arXiv:2305.19555*, 2023.

[56] S. Ott, K. Hebenstreit, V. Liévin, C. E. Hother, M. Moradi, M. Mayrhauser, R. Praas, O. Winther, and M. Samwald, "ThoughtSource: A central hub for large language model reasoning data," *arXiv preprint arXiv:2301.11596*, 2023.

[57] Z. Tao, Z. Jin, X. Bai, H. Zhao, Y. Feng, J. Li, and W. Hu, "EvEval: A Comprehensive Evaluation of Event Semantics for Large Language Models," *arXiv preprint arXiv:2305.15268*, 2023.

[58] N. Riccardi, and R. H. Desai, "The Two Word Test: A Semantic Benchmark for Large Language Models," *arXiv preprint arXiv:2306.04610*, 2023.

[59] N. Lee, N. M. An, and J. Thorne, "Can Large Language Models Infer and Disagree Like Humans?," *arXiv preprint arXiv:2305.13788*, 2023.

[60] Y. Liu, T. Han, S. Ma, J. Zhang, Y. Yang, J. Tian, H. He, A. Li, M. He, and Z. Liu, "Summary of chatgpt-related research and perspective towards the future of large language models," *Meta-Radiology*, pp. 100017, 2023.

[61] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, and A. Kasirzadeh, "Ethical and social risks of harm from language models," *arXiv preprint arXiv:2112.04359*, 2021.

[62] R. T. McCoy, E. Pavlick, and T. Linzen, "Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference," *arXiv preprint arXiv:1902.01007*, 2019.

[63] B. Wang, W. Chen, H. Pei, C. Xie, M. Kang, C. Zhang, C. Xu, Z. Xiong, R. Dutta, and R. Schaeffer, "DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models," *arXiv preprint arXiv:2306.11698*, 2023.

[64] V. Rawte, A. Sheth, and A. Das, "A survey of hallucination in large foundation models," *arXiv preprint arXiv:2309.05922*, 2023.

[65] K. Ahuja, R. Hada, M. Ochieng, P. Jain, H. Diddee, S. Maina, T. Ganu, S. Segal, M. Axmed, and K. Bali, "Mega: Multilingual evaluation of generative ai," a*rXiv preprint arXiv:2303.12528*, 2023.

References

[66] V. D. Lai, N. T. Ngo, A. P. B. Veyseh, H. Man, F. Dernoncourt, T. Bui, and T. H. Nguyen, "Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning," *arXiv preprint arXiv:2304.05613*, 2023.

[67] Y. Zhao, T. Pang, C. Du, X. Yang, C. Li, N.-M. Cheung, and M. Lin, "On evaluating adversarial robustness of large vision-language models," *arXiv preprint arXiv:2305.16934*, 2023.

[68] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, and E. Brunskill, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[69] J. Broughel, "OpenAI Is Now Unambiguously Profit-Driven, And That's A Good Thing," https://www.forbes.com/sites/jamesbroughel/2023/12/09/openai-is-now-unambiguously-profit-driven-and-thats-a-good-thing/, 2023.

[70] IMF, *ARTIFICIAL INTELLIGENCE*: *What AI means for economics*, International Monetary Fund, 2023.

[71] Huggingface, "SupportingOpenSourceandOpen ScienceintheEUAIAct," 2023.

[72] X. Liu, G. Wang, and D. Zha, "Fingpt: Democratizing internet-scale data for financial large language models," *arXiv preprint arXiv:2307.10485*, 2023.

[73] N. Vigdor, "Apple card investigated after gender discrimination complaints," T*he New York Times*, vol. 10, 2019.

[74] ISO, "ISO/IEC TS 5723:2022 Trustworthiness Vocabulary," International Organization for Standardization, 2022.

[75] FTC, *Gramm-Leach-Bliley Act*, Federal Trade Commission, 2023.

[76] Financial Conduct Authority, "Call for Input: Potential competition impacts from the data asymmetry between Big Tech and firms in financial services," https://www.fca.org.uk/publications/calls-input/potential-competition-impacts-data-asymmetry-big-tech-firms-financial-services, 2023.

[77] OWASP, *LLM AI Cybersecurity & Governance Checklist*, 2024.

[78] CRFM. "The Foundation Model Transparency Index," https://crfm.stanford.edu/fmti/.

[79] Slido. "The easiest way to make your meetings interactive," https://www.slido.com/.

[80] J. Weizenbaum, "ELIZA—a computer program for the study of natural language communication between man and machine," *Communications of the ACM*, vol. 9, no. 1, pp. 36-45, 1966.

[81] E. D. Liddy, "Natural language processing," 2001.

[82] C. Jones, and B. Bergen, "Does GPT-4 Pass the Turing Test?," *arXiv preprint arXiv:2310.20216*, 2023.

[83] X. Liu, and W. B. Croft, "Statistical language modeling for information retrieval," *Annu. Rev. Inf. Sci. Technol.*, vol. 39, no. 1, pp. 1-31, 2005.

[84] B.-H. Juang, and L. R. Rabiner, "Automatic speech recognition–a brief history of the technology development," *Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara,* vol. 1, pp. 67, 2005.

References

[85] P. Azunre, *Transfer learning for natural language processing:* Simon and Schuster, 2021.

[86] A. Kovačević, and D. Kečo, "Bidirectional LSTM networks for abstractive text summarization." pp. 281-293.

[87] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, and K. Macherey, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[88] R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang, and T. Liu, "On layer normalization in the transformer architecture." pp. 10524-10533.

[89] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, and Z. Dong, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[90] E. Roth, "Apple's kind of a bank now / It'll handle the financial side of its new Pay Later service," https://www.theverge.com/2022/6/8/23160400/apple-kind-of-bank-now-pay-later-bnpl-financial-services, 2022.

[91] Bank of England, "The AI Public-Private Forum: Final report," https://www.bankofengland.co.uk/research/fintech/ai-public-private-forum, 2022.

[92] Bank of England and Financial Conduct Authority, "FS2/23 – Artificial Intelligence and Machine Learning," 2023.

[93] Capco, "Artificial Intelligence," *The Capco Institute Journal of Financial Transformation*, 2023.

[94] OECD, "Artificial Intelligence, Machine Learning and Big Data in Finance Opportunities, Challenges and Implications for Policy Makers," https://www.oecd.org/finance/financial-markets/Artificial-intelligence-machine-learning-big-data-in-finance.pdf.

Future Work from the Turing Community

Building on the insights revealed in this study, the Alan Turing Institute plans to collaborate with partners to investigate funding opportunities to develop the identified focus areas for ensuring trustworthy and safe integration of large language models.

Further, current opportunities to watch for include:

- The Alan Turing Institute plans to execute a red-teaming event specifically targeting financial LLMs to facilitate stress testing from interdisciplinary groups, inform confidence levels and guide opportunities to strengthen integrity and resilience for open-weight and open-source models, and build on recent read-teaming activities [2, 3]
- The Alan Turing Institute is organising a workshop on AI in the Digital Economy with an emphasis on presenting state-of-the-art research from the academic community during the upcoming International Conference on AI and the Digital Economy 2024 [4]
- The Turing Synthetic Data Interest Group [5] will be organising an event to explore the opportunities and challenges.

turing.ac.uk
@turinginst