ESMA
European Securities and Markets Authority

Institut Louis Bachelier

The Alan Turing Institute

LEVERAGING LARGE LANGUAGE MODELS IN FINANCE:

Pathways to Responsible Adoption

This report is part of a collaboration between the European Securities and Markets Authority (ESMA), the FaiR (Finance and Insurance Reloaded) programme at Institut Louis Bachelier and the FAIR (Framework for Responsible Adoption of Artificial Intelligence in the Financial Services Industry) programme of the Alan Turing Institute. The views expressed in this publication are privately held by the authors and should not be attributed to ESMA, Institut Louis Bachelier or the Alan Turing Institute.

AUTHORS:

Bagattini Giulio (ESMA)

Brière Marie (Institut Louis Bachelier)

Guagliano Claudia (ESMA)

Maple Carsten (The Alan Turing Institute)

Sabuncuoglu Alpay (The Alan Turing Institute)

WORKSHOP PARTICIPANTS:

Balague Christine - Institut Mines Telecom

Bertucci Louis - Institut Louis Bachelier

Bonaita Alessandro - Generali

Chen Yujia - University of Edinburgh

Choné Anne - ESMA

Dymacz Aleksandra - OECD

Elie Romuald - Deepmind

Fliche Olivier - ACPR

Gomez Teijeiro Lucia - Geneva University

Gourier Elise - ESSEC

Hammouda Maysara - Predictiva

Ionita Laura - ESMA

Krasniqi Dafnis - Institut Louis Bachelier

Lauridsen Nico - EUI

Lefort Baptiste - AI for Alpha

Leote De Carvalho Raul - BNP PAM

Lucas Iris - AMF

Ly Antoine – SCOR

Marshall William - ESMA

Martinez Luis - ESMA

Masson Corentin - AMF

Mery Rami - Amundi

Ohana Jean Jacques - AI for Alpha

Otaegui Alain - EBA

Papiotis Sotiris – ESMA

Piazza Federico - ESMA

Picault Matthieu - University of Orléans

Rava Matteo – ESMA

Ritola Tuomas - FIN-FSA

Rouil Guillaume - AXA IM

Smits Artis - CB Latvia

Szpruch Lukasz - The Alan Turing Institute

Townson Sian - OliverWyman

Special thanks to **Louis Bertucci**, **Anne Choné**, **Dafnis Krasniqi**, **Corentin Masson** and **Sian Townson** for useful comments and contributions to the report.

CONTENTS

ABSTRACT

This report presents the summary of discussions held during a workshop on the use of large language models (LLMs) in the financial industry organised in June 2024 by the European Securities and Markets Authority, the Alan Turing Institute and the Institut Louis Bachelier. The workshop engaged 38 technology and finance experts to discuss three main issues around (1) the current use of LLMs and their potential applications in the financial industry, (2) the risks and challenges associated with their use, and (3) the steps necessary for ensuring their responsible adoption.

Generative LLMs are increasingly used in the financial industry to achieve operational efficiencies in tasks involving text analysis and production, but they are also increasingly deployed for public communication and customer interaction. This raises potential issues, often tied to legal, ethical and reputational harm. Against this backdrop, many financial organisations are developing pathways to responsible LLM adoption that deal with the topics of model robustness, data dependency, security and privacy, fairness and accountability.

The finance sector can benefit from the establishment of appropriate evaluation metrics for the use of LLMs, including benchmarks, and the development of industry standards. An appropriate supervisory framework, together with adequate staff training, can facilitate this effort. Meanwhile, the downside posed by the carbon footprint of LLMs may need to be carefully evaluated as the technology spreads and its use becomes integral to the everyday operations of global businesses in the finance sector and beyond.

1 . Introduction

1 . Introduction

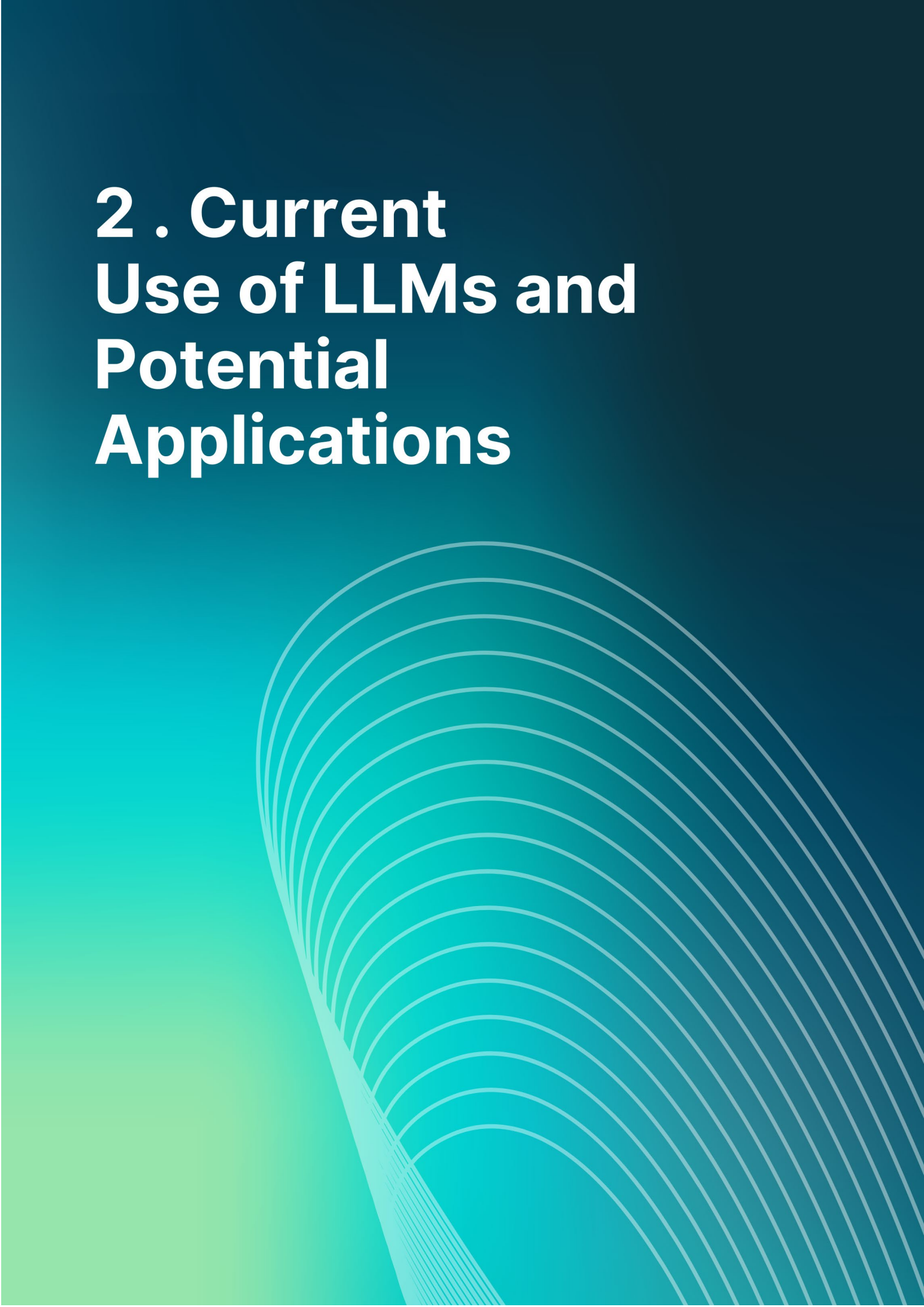
Historically, the finance industry has been an early adopter of many technological advancements, from electronic systems to Big Data and artificial intelligence (AI). However, the adoption of new technologies is often cautious, due to risk aversion, regulatory compliance and legacy systems. Regulatory bodies set standards and guidelines to protect customer interests and create guardrails, while aiming not to stifle innovation. Financial institutions typically conduct thorough due diligence and rigorous testing before implementing new technologies to mitigate potential damage. The resulting configuration, while potentially conservative, aims to balance innovation and risk to meet evolving business needs while safeguarding the financial system's stability.

Large language models (LLMs), and particularly their generative versions, have found several application areas in the financial services industry. The ability of these models to process and analyse vast amounts of text can provide significant improvements in efficiency, accuracy, and decision-making. From sentiment analysis and risk assessment to generating investment insights and personalised financial advice, LLMs have the potential to transform how financial institutions operate. At the same time, as with any emerging technology, a fast uptake in the use of LLMs may come with different issues and challenges, including robustness, security, fairness, and regulatory compliance. In a complex and highly interconnected sector, it is thus essential to assess the trade-offs between the benefits and the potential side effects generated by the use of these models and evaluate measures to monitor and control the most material risks.

The workshop held in June 2024 aimed to identify opportunities and risks associated with LLMs by leveraging existing knowledge and expertise. Jointly organised by the Finance and Insurance Reloaded (FaIR) programme of Institut Louis Bachelier, the FAIR (Framework for Responsible Adoption of AI in the Financial Services Industry) programme of the Alan Turing Institute and the European Securities and Markets Authority (ESMA), the workshop brought together around 40 leading experts in the field, including data and technology managers from major financial institutions, financial sector consultants, researchers and regulators. The participants examined the current and emerging opportunities for using LLMs in the financial services sector, explored the main risks and challenges associated with their use, and articulated reflections around pathways for their responsible and trustworthy adoption. Starting from the financial industry's existing and emerging LLM-related policies – both for internal use and for customer-facing applications – the discussions reflected on the robustness, security, fairness and integrity of LLMs.

This report builds on the considerations and insights that emerged during the workshop. Section 2 gives an overview of the current use of LLMs and their nascent applications in the finance industry. Section 3 presents the potential harms associated with their use. Section 4 offers reflections on the possible steps for facilitating their trustworthy adoption. Finally, Section 5 provides context on future prospects and the regulatory framework.

2 . Current Use of LLMs and Potential Applications



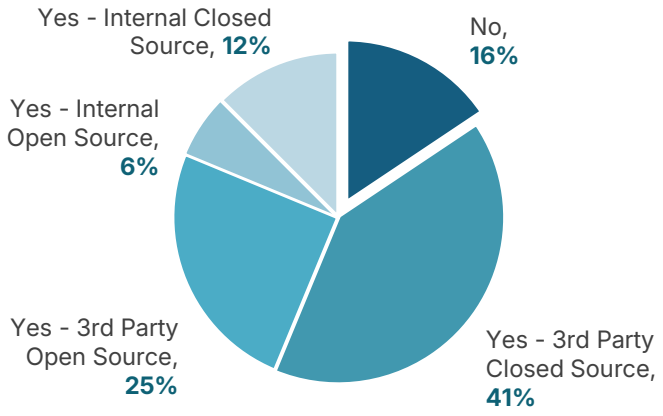
2 . Current Use of LLMs and Potential Applications

The workshop discussed the use of generative LLMs in the finance industry for both internal and customer-facing applications, examining both the current use of LLMs and the potential future developments that participants foresee in their organisations.

Participants were asked the following question.

Workshop Question 1:

Has your organisation deployed LLM-powered services to support daily work-related tasks? If yes, what type of LLMs are in use?



Note: "Closed source" means that the model's code, training data, and architecture are proprietary and not publicly available – examples include OpenAI's GPT-4 and Google's Gemini. "Open source" means that the model's code, weights and, sometimes, training data are publicly available, allowing anyone to use, modify, or improve it – examples include Meta's LLaMA (as regards the model's weights) and Allen AI's OLMo (for both model weights and training data). "Third party" means that the LLM is deployed by users without any modifications. "Internal" means that the LLM is developed internally by a firm or – more commonly – customised before deployment, for example by fine-tuning with internal data, safeguarding, instruction-tuning, etc. Number of respondents: 32.

A large majority of respondents **(85%)** indicated that their organisation already uses LLMs. Out of these, almost half use LLMs provided by third parties and developed on a closed-source model. Less than **20%** of the respondents stated the use of internally customised LLMs, either open- or closed-source.

Financial Services Application Areas

LLMs trained on vast financial datasets can serve as virtual assistants for finance professionals to more efficiently perform tasks such as data extraction, summarisation, and coding. This can significantly reduce the time and effort required for manual data processing, allowing workers to focus on more strategic and value-added activities. Hence, many expect that the progressive integration of these systems into daily office tasks will yield significant **productivity gains**. Alongside these support functions, LLMs have the potential to conduct further, more autonomous tasks. Applications of LLMs that are emerging and are being tested in financial practice include:

- **Automating compliance processes** by analysing regulatory documents and identifying relevant clauses, pulling together data from disparate sources to compile draft reports.
- Generating **automatic reporting, translations and summarisation** of documents, calls or meeting minutes for internal use or for clients.
- Being incorporated into customer service platforms to handle inquiries and provide **immediate support to customers**.

- Powering interactive **financial chatbots** that provide tailored advice and recommendations to individuals based on their financial goals, risk tolerance, and current market conditions.[1]
- Analysing **market sentiment** and identify emerging trends by processing news articles, social media data, and other unstructured text sources.
- Analysing historical financial information to **make predictions** about future market trends and economic conditions, thereby helping financial institutions in risk assessment, portfolio optimisation, and investment strategies.
- Assisting in **ESG analysis and reporting** by extracting relevant data from various sources, analysing the company's ESG performance, and generating sustainability reports that address stakeholders' demands for sustainable and responsible investing.

Next, participants answered the following open-ended question.

Workshop Question 2:

What are the most promising opportunities for integrating LLMs in financial decision-making processes?

The answers included the following application areas (from most to least frequently mentioned):

- **Document summary and management**, coverage review expansion and productivity improvement, especially for unstructured data.
- Improving the overall **research phase** before financial decision-making, by making large batches of documents easily analysable and highlighting all the relevant information.
- **Analysing vast volumes of information** quickly and efficiently (including numeric figures, also legal texts, images, etc).
- Analysing tasks and **translating requests into code**, in combination with other AI algorithms.
- Content intelligence and **content generation** for multiple use cases (HR, audit, legal, customer service).
- **Accessing large databases** and providing more comprehensive and precise guidance and information.
- **Novel services for customers** and enhanced competition and innovation within the offer of financial services by combining AI tools with Open Finance.
- **More tailored proposals for customers**, aligned with their needs and profile, thus also contributing to improve financial inclusion and literacy.

In these applications, most participants emphasised that LLMs are **augmenting the capabilities of experts** rather than replacing them.

Participants also emphasised that their organisations were focusing on the use of LLMs where the **synthesis and analysis of large amounts of text can improve the efficiency** of the product, system or process. This includes various tasks: classification of a text into subtopics (for example to analyse firms' communication or regulatory filings), summarisation of reports into a shorter or less complex text that can be disseminated to clients, and search of precise information (for example, to check compliance with legal obligations).

Use cases were still **mostly 'internal'**, with no 'direct' interaction between the LLM and the customer. One exception to this common arrangement might be robo-advice, although the role of

[1] In this field, competition enabled by LLMs may actually represent a challenge for financial intermediaries. An ecosystem of financial investing tools powered by AI has bloomed especially outside the perimeter of regulated financial institutions. Retail investors could also turn to general-purpose LLMs for financial advice. However, ESMA (2025) warned that this practice involves risks. Poorly vetted AI tools can generate inaccurate or misleading advice that may result in financial losses.

LLMs in these systems appeared to be still limited to basic interactions such as requesting data inputs from the user, rather than providing financial advice on the client's request.

Considering all these use cases, four main potential application areas of LLMs can be delineated as follows.

(1) Public communication and customer engagement. It includes functions such as financial communication to the public (e.g. to simplify technical jargon) and customer service. In marketing and customer service, LLM-based chatbots are already used to improve customer experience and customer conversion rates.

(2) Financial services safety. It encompasses various functions, 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 identify patterns and anomalies that indicate fraudulent activities or financial crimes.[2]

(3) Financial insight generation. It includes functions such as market surveillance and the generation of insights related to markets, business finance data, personal investment and ESG (environmental, social and governance) assessments.

(4) Financing and investment activities. It includes functions such as asset management, investment banking, treasury optimisation, private equity and venture capital strategy development.

PUBLIC COMMUNICATION AND CUSTOMER ENGAGEMENT	FINANCIAL SERVICES SAFETY	FINANCIAL INSIGHT GENERATION	FINANCING AND INVESTMENT ACTIVITIES
• Financial communication • Customer service and chatbots • Investing apps	• Detecting and preventing fraud • Risk assessment • Market surveillance	• Generation of market insights and reports • Data analysis • Coding assistant	• Asset management • Loan financing • Investment banking • Private equity and venture capital strategy
Examples: • Answer questions from clients • Provide personalised investment reports • Generate automatic marketing reports from data or graphs • Generate automatic translation	**Examples:** • Analyse social media • Summarise large amount of textual information (e.g. contracts) • Flag inconsistencies and behavioural anomalies	**Examples:** • Financial sentiment analysis • Entity recognition • News search • Classification • Hypotheses formulation	**Examples:** • Automatic extraction of information from documents (e.g. bank statements, tax forms) • Analysis of firms' communication (e.g. regulatory fillings, patent information) to generate buy or sell signals to be acted upon in the investment process

Timeliness

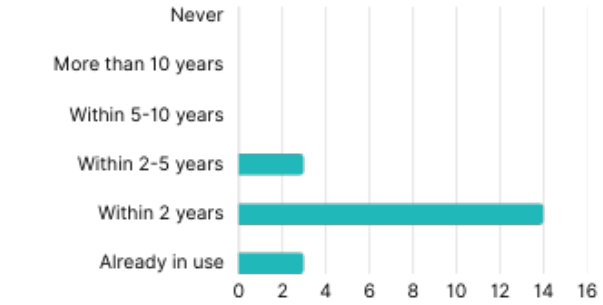
Workshop participants were asked to express their opinion about the most likely timeline for their organisations to adopt LLMs for each of the application areas listed above.

[2] For example, see https://medium.com/slope-stories/slope-transformer-the-first-llm-trained-to-understand-the-language-of-banks-88adbb6c8da9.

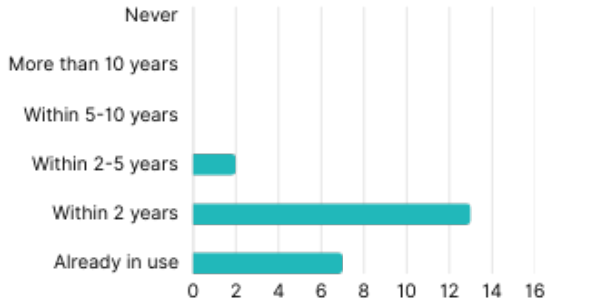
Workshop Question 3:

When do you foresee the earliest integration of LLMs in <given application category>?

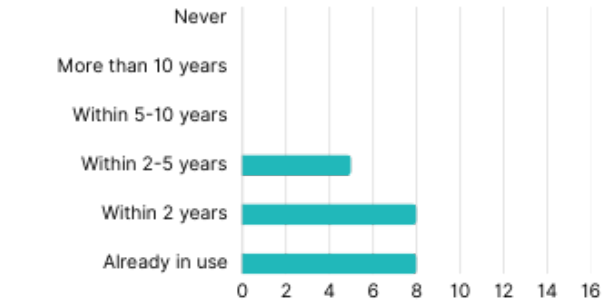
When do you foresee the earliest integration of LLMs in public communication and engagement services (e.g. financial communication, customer interaction)?



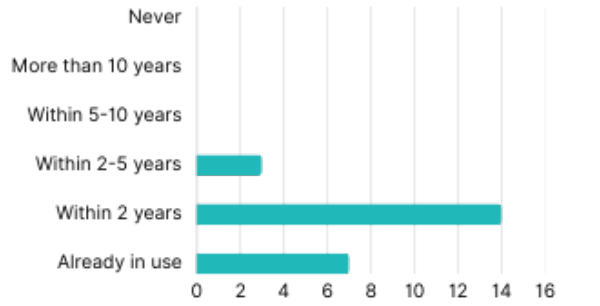
When do you foresee the earliest integration of LLMs in financial insight generation services (e.g. market and trade surveillance, personal investment reports)?



When do you foresee the earliest integration of LLMs in the "financing and investment activities" (e.g. loan financing, investment, portfolio or capital allocation decisions)?



When do you foresee the earliest integration of LLMs in financial services safety (e.g. fraud detection, risk management)?



The survey results show that workshop participants were clearly upbeat as to the prospects for their organisations to leverage LLMs, with most respondents either foreseeing a relatively rapid adoption of this technology (within 2 years) in their organisation or stating that it is already in use across all the identified application areas. Financing and investment activities is the area where most respondents (8 out of 21, or 38%) indicated their organisation was already using LLMs, although some respondents (5 out of 21, or 24%) did not foresee a rapid adoption within 2 years. Public communication and customer engagement displayed the lowest rate of adoption among all the four areas, with few organisations (3 out of 20, or 15%) alleged to have already deployed LLMs in related functions.

This somewhat sets apart customer-facing functions from other application areas in the experimentation and adoption of LLMs. The former may be considered more sensitive due to their more direct reputational and compliance risks and thus require a higher degree of maturity and more comprehensive testing of new technologies before they are deployed.

Nevertheless, notably, most of the respondents (14 out of 20, or 70%) believed that their organisations would make up for this lag relatively quickly, starting to integrate LLMs into these functions within 2 years.

3. Potential Harms from the Use of LLMs

3. Potential Harms from the Use of LLMs

Risks from integrating AI systems deep into financial services businesses can emerge at various levels, including data, models, and governance. In complex interconnected systems, these risks can propagate and compound, with the potential to become systemic and affect financial stability.[3] The workshop discussion focused on assessing harms from LLMs directly impacting the financial institutions which deploy them and their clients.

Financial institutions, bound by comprehensive regulatory requirements and growing concerns around reputational risk, may face constraints in deploying AI systems that lack explainability or fail to deliver outputs predictably, consistently, and without significant risk of error.

The workshop participants were presented with the following categories of potential harms stemming from the integration of LLMs into financial services.

- **Human-rights harms**. LLMs may violate privacy rights by mishandling sensitive financial data or generating outputs that compromise confidentiality. The use of AI in decision-making could lead to unfair treatment, undermining individuals' rights to fair access to financial services.
- **Well-being harms**. Misleading financial guidance from LLMs can cause financial distress, anxiety, or economic hardship for clients. Employees relying on AI-generated outputs may experience job insecurity or stress due to automation-driven changes in the workplace.
- **Representational harms**. Biases embedded in training data can result in discriminatory financial outcomes, such as denying loans to marginalized groups. This can reinforce systemic inequalities, expose institutions to regulatory scrutiny, and damage client relationships.
- **Quality of service harms**. LLMs may produce inconsistent or incorrect financial insights, impacting decision-making for both institutions and clients. Poorly trained models could lead to subpar customer service, miscommunication, or inadequate risk assessments, reducing trust in financial services.
- **Legal and reputational harms**. LLMs may generate misleading or inaccurate financial advice, leading to legal liabilities and regulatory violations. Financial institutions risk lawsuits, fines, and reputational damage if clients suffer losses due to AI-generated misinformation or bias in decision-making.

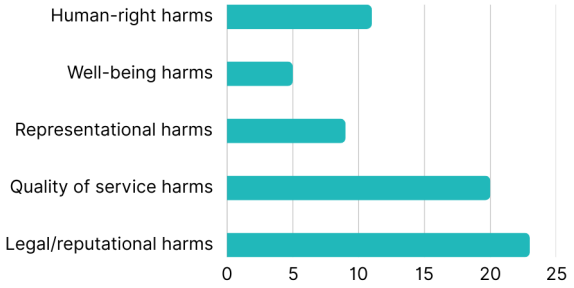
To gauge their perception of these risks, participants were asked to assess the potential harms from the use of LLMs in the different functional areas.

[3] For considerations on the financial stability risks of AI, such as third-party dependencies and correlated risks, see Financial Stability Board (2024), The Financial Stability Implications of Artificial Intelligence.

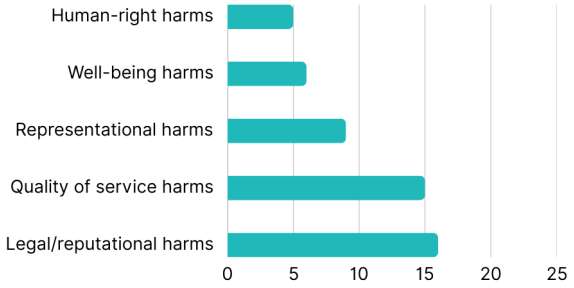
Workshop Question 4:

Which of the following would you consider as potential harms of integrating LLMs into <given application category>?

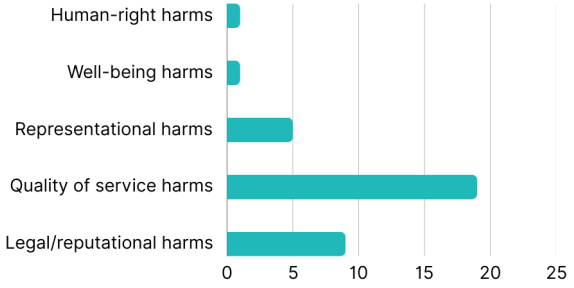
Which of the following would you consider as potential harms of integrating LLMs in public communication and engagement services (e.g financial communication, customer interaction)?



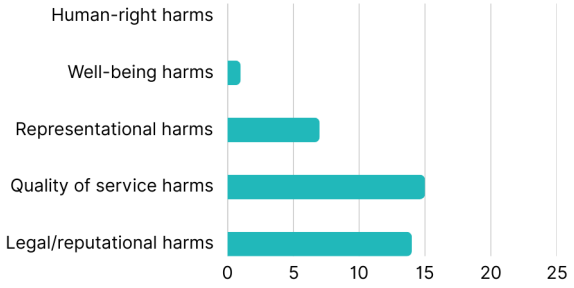
Which of the following would you consider as potential harms of integrating LLMs in financial services safety (e.g fraud detecion, risk management)?



Which of the following would you consider as potential harms of integrating LLMs in financing and investment activities (e.g hypothesis generation, data analysis, investment strategy construction)?



Which of the following would you consider as potential harms of integrating LLMs in financial insight generation services (e.g market and trade surveillance, personal investment reports)?



Across the four applications areas, harms related to the quality of service and legal and reputational harms were considered those with the highest risk of materialising among the proposed categories. Public communication and engagement services was the area where most respondents indicated potential harm, corroborating the hypothesis that the increased caution in integrating LLMs into these functions be tied to the inherently more prominent reputational and compliance risks.

Concerns around the quality of service were frequent also in relation to financing and investment activities, possibly because applications in this category (for instance, determining investment strategies) are considered more financially sensitive to mishaps in modelling and output.

Human rights and well-being harms, which are often mentioned as potential societal issues related to broader AI advancements, did not feature prominently among participants' answers (with the partial exception of public communication and engagement), suggesting that these themes may be considered less material in the context at hand.

4. Towards Responsible Adoption

4. Towards Responsible Adoption

The next stage of the workshop aimed to address the sources of the harms laid out above and tabled a discussion around the design of strategies to mitigate these risks. The workshop discussion was structured around five main issues related to the adoption of LLMs: (**1) robustness; (2) data dependency and asymmetry; (3) security and privacy; (4) fairness and bias; and (5) accountability and explainability**. These topics are largely aligned with the seven ethical principles for AI laid out in the 2019 Ethics guidelines[4] for trustworthy AI of the AI HLEG (European Commission High Level Expert Group on AI) appointed by the European Commission, namely: human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; accountability; societal and environmental well-being.[5]

The discussion also touched upon the carbon footprint of LLMs, in line with the last principle of the AI HLEG, and evaluated the necessary skills and staff training that an increasing adoption of LLMs will require.

Robustness

Robustness for an AI system is often defined as **producing accurate and reliable outputs and performing as expected under various conditions**. In the case of LLMs, this means that the model would be expected to produce the same or very similar answers to similar prompts over time, ensuring its output can be relied upon and reproduced under a variety of conditions, including unexpected inputs, adversarial examples, and changes in the environment. In this sense, robustness can be represented through two dimensions: reliability and reproducibility. The following questions arise:

- **Reliability and consistency:** Does the LLM answer the question/handle the topic/solve the task properly? Does it refuse to answer? What is the model's ability to provide consistent outputs for similar inputs? If the same question is asked multiple times, a robust LLM should produce similar answers – barring any changes in the provided context or re-training.
- **Generalisation:** A robust LLM should generalise well to new, unseen data that it was not explicitly trained on.
- **Noisy or adversarial input**: In case of input perturbations (for example, noisy, incomplete or ambiguous input, including typos, slang or informal language), or adversarial input (carefully crafted input designed to confuse the model), what is the model's behaviour?
- **Error handling:** When the model responds incorrectly, what is the level of inaccuracy? Are there metrics to quantify it? Does the LLM have mechanisms to recognise errors?
- **Performance under stress:** Robust LLMs should be able to maintain performance even under high load or when processing large volumes of requests simultaneously. This is particularly important for applications in production environments.

Practical Challenges

- A major challenge is measuring the accuracy of **general-purpose LLMs while keeping multiple business use-cases in mind.**[6] Participants stated that organisations tend to **use**

[4] See https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[5] See the Ethics guidelines for trustworthy AI. The EU AI Act (see Section 5) recalls the validity of these principles as a basis for the drafting of codes of conduct under the Regulation (EU AI Act, Recital 27).

[6] General-purpose LLMs, such as GPT-4, are pre-trained on vast and diverse datasets. These models are designed to handle a wide range of tasks, from text generation to question answering, without requiring further training or customisation.

heuristics such as comparing the outputs of various models, making random consistency checks between input text and the models' output, and using different LLMs to check and validate each other's results.

- Achieving a satisfactory degree of robustness generally requires **different assessment methodologies** based on the use case, the nature of the task at hand and the scope. For example, in data-driven applications that require specific answers based on a specific text (or data), participants expect high reliability from the model. Furthermore, robustness in the generative AI (GenAI) space is difficult to define objectively because of the inherent challenges of quantifying divergences between the generated outputs. Some participants pointed out that there are **substantial differences between the models** in terms of outputs.
- GenAI tools are **inherently creative** and developed mostly for general-purpose capabilities to assist the user in a wide variety of applications, which makes robustness testing even more challenging. **Some deviations of an LLM's output from what would be considered a correct answer are not "explainable"**.[7] Sometimes no patterns emerge allowing to explain "hallucinations" – i.e., the generation of factually false or misleading information – which is central to ensuring robustness. LLMs are inherently built to give "confident" answers, which makes it harder for users to identify false or misleading content.
- When "fine-tuning" a model, it is challenging to strictly constrain it to a specific field of activity, which can lead to imprecise or irrelevant answers.[8] Assessing consistency is challenging due to the dynamic nature of fine-tuning, which can alter a model's behaviour due to the vast number of parameters. This makes exact reproducibility difficult.
- LLMs are **bound to human-produced knowledge** and can learn false information from the internet (the source of the bulk of their data), which can harm their performance and reliability.
- Proprietary, closed-source models often **lack transparency**, which complicates reproducibility since the same query might yield different results over time, as proprietary models could be trained with new data in the meantime. Conversely, open-source models offer relatively more control, but they might not perform as well as proprietary models.
- All these issues become especially significant when practitioners are left without guidelines from model providers and internal IT teams on how to best make use of the tools. Many participants pointed to the **importance of being trained** in the appropriate use of LLMs, their functioning and their potential pitfalls.

Possible Mitigation Approaches

- **Fine-tuning, retrieval-augmented generation (RAG), and fact-checking mechanisms** can help mitigate hallucinations.[9]
- However, users need to be aware that, depending on the model's architecture and training, it may be impossible to entirely prevent an **LLMs from "hallucinating"**. Being this a deliberately creative tool, its management should be centered around making sure that a certain rate of wrong or inaccurate output does not compromise the tool's added value , or there is a mechanism to spot the undesired response before it goes further down the pipeline. **Human feedback** is an important component of making models more robust.
- The impact of output inaccuracies varies by application and can be of different nature, presenting firms with different trade-offs, which should be subject to a **global cost-benefit analysis**. For instance, hallucinations in an LLM that underpins a customer service chatbot might lead to reputational damage for the firm, while in the case of an LLM that assists an investment research analyst in retrieving summaries of company reports, it might increase the risk of inefficient decision-making and financial losses.

[7] In AI, explainability broadly refers to the ability to describe an AI model's internal workings or outcomes in understandable terms.
[8] Fine-tuning is a technique for customising and optimising the performance of LLMs for specific use cases. By further training a pre-trained LLM on a labeled dataset related to a particular task, fine-tuning can improve the model's performance.
[9] Retrieval-Augmented Generation (RAG) enhances language models by integrating an external retrieval mechanism. Instead of relying solely on pre-trained knowledge, RAG fetches relevant documents or facts from an external database (e.g., Wikipedia, vector search) and uses them to generate more accurate and contextually grounded responses.

- One way to **reduce errors is to ask LLMs precisely where to look for information**. LLMs can be asked to generate both **reasoning traces and task specific actions**. For example, an LLMs can be asked to answer a query with thought (such as providing a reasoning chain), action (such as searching information on the web, from specific sources, over a certain period) and observation steps. Some providers have embedded this into their models' workings via specific "search" and "reason" options that complement the user's prompt.[10]
- A metric for robustness could include the **rate of hallucinations** obtained under different testing conditions. Some participants use **external evaluation tools, such as Giskard, that utilise adversarial attacks**, and model testing for efficiency under various conditions.[11] However, there is no baseline used by the entire industry to determine accuracy. Adopting a set of standard key performance indicators for robustness may facilitate the assessment of different models. For example, **LLM benchmarks** adapted for specific financial tasks can be a starting point to improve the robustness evaluation process.[12]
- **LLMs can also be used to evaluate each other**, for example, to "red-team" each other (i.e. simulating attacks) for out-of-sample results and identify areas where an LLM seems to be inferior.
- All these factors notwithstanding, it is important to acknowledge that LLMs are not optimised for accuracy in the same way as other more traditional, non-generative AI models are; thus, the same standards and metrics around robustness of certain machine learning models cannot generally be translated to LLMs. Adopters in the financial industry should thus take a more adapted approach and set of expectations around unpredictable behaviour.

Data Dependency and Asymmetry

Data dependency and asymmetry challenges in the financial industry have been relevant since long before LLMs came to prominence, especially since the rise of Big Data and AI use for financial services.[13] LLMs are no exception, as they rely on datasets that dominant incumbent firms are likely to have an advantage in obtaining.

Practical Challenges

- Data dependency and asymmetry in LLM development and fine-tuning can impair **transparency and market competition**.
- The fixed costs of training and fine-tuning an LLM are much larger than the variable costs from its use. To train an LLM on a specific task, a **large amount of data is needed**, which can be very costly to label. Financial data itself can be costly to obtain. While historical data may be accessible, real-time data is often very expensive. Large players typically have more resources and access to data, while smaller entrants may face challenges from limited data availability.
- Training and deploying models may **require significant computing power and technical skills**. Large companies have more resources to overcome these financial and technical challenges than smaller companies. Smaller financial institutions may struggle to acquire

[10] See, for instance, https://openai.com/index/introducing-chatgpt-search/ and https://www.thedailystar.net/tech-startup/news/chatgpt-now-has-reason-button-similar-deepseek-3814221.

[11] In AI, adversarial attacks are inputs that trigger the model to produce undesired or incorrect output. Giskard is an open-source Python library that detects performance, bias and security issues in LLMs and other AI models, including hallucinations, harmful content generation, robustness issues, sensitive information disclosure, stereotypes and discrimination.

[12] An LLM benchmark is a standardised performance test used to evaluate various capabilities of AI language models. A benchmark usually consists of a dataset (with expected answers), a collection of questions or tasks, and a scoring mechanism. See, for example, HuggingFace's Big Benchmarks Collection for a comprehensive list of benchmarks. Further methods to evaluate the quality of LLMs' output include the "BLEU" method – which judges the quality of a machine translation by measuring its closeness to a set of reference human translations – and the "ROUGE" method – which measures "recall", i.e. how much content from one or more references is actually contained in the LLM's output.

[13] Data asymmetry refers to the imbalance where organisations with more access to high-quality, proprietary financial data gain a competitive edge in model performance and insights, leaving others with less data at a disadvantage.

staff with sufficient theoretical understanding and practical skills to effectively deploy LLMs.
- Data asymmetry between companies also exists from the perspective of the data used to train models. There is a **large-firm data hegemony**, in the sense that data on large firms is easily retrievable, while small firms do not make as much data readily available for models to use, especially in the EU.

Possible Mitigation Approaches

- **Synthetic data** can be used in financial organisations to supplement or replace real data, helping overcome data limitations, protect privacy, and reduce biases. This allows for model training, testing, and validation without relying on sensitive or proprietary data, improving model performance while mitigating data asymmetry issues. Although these methods already have recognition in finance, some workshop participants questioned the effectiveness and utility of synthetic data. They emphasised that their effectiveness depends on the business case. Privacy, utility, and fidelity appeared as the key terms when considering synthetic data and their application.
- Practitioners **need centralised testing databases to audit models** and to test the "usefulness" of synthetic data. For this, creating a taxonomy and applying a testing methodology through a structured process can be beneficial.
- **Regulations around data anonymisation and access** could help mitigate asymmetry. Proposals include creating systems akin to open banking, allowing controlled data access for various actors. This is one of the objectives of the UK open banking system. In the EU, several important initiatives aim to facilitate access to financial data by a large audience and to address data asymmetry issues between larger and smaller firms. By 2027, the **European single access point (ESAP)** will provide easy access – in extractable and machine-readable format – to centralised public information about companies and investment products – such as financial statements, prospectuses and ESG information – throughout the EU.[14] **The EU data hub**, currently under deployment, will allow market participants to use (synthetic) supervisory data upon request, for example for AI model testing.[15] The financial data access (FIDA) proposal builds on the open banking framework with a view to facilitating the sharing and re-use of customer data (upon consent) for a wide range of financial sectors and products.[16]

Security and Privacy

Trustworthy AI systems need to withstand unexpected events, adversarial attacks, and evolving conditions. **Security** focuses on protecting against unauthorised access and use, while **resilience** emphasises the ability to maintain functionality and recover from disruptions. These systems should also be designed to **protect individual privacy** throughout their lifecycle. This includes minimising data collection, ensuring secure data storage, and implementing privacy-preserving techniques to safeguard personal information.

Practical Challenges

- Participants highlighted **cyber security risks** as the key concern (several of them mentioned this as a limiting factor for GenAI adoption). While technical mitigants were being considered and developed, human error remained an inherently weak point, just like in all

[14] The ESAP is a flagship action of the EU Capital Markets Union (CMU) Action Plan and a concrete realisation of the EU Digital Finance Strategy. Regulation (EU) 2023/2859 , which entered into force in January 2024, provides that ESMA is due to establish and operate ESAP by 10 July 2027.,
[15] The EU digital finance platform's data hub is part of the 2020 European strategy for data, through which the EU committed to boosting the development of trustworthy data-sharing systems..
[16] The FIDA proposal, which is part of the EU Digital Finance strategy, seeks to establish a framework governing access to and use of customer data in finance (sometimes also referred to as the open finance framework) with the objective of simplifying the access to customers' financial data.

IT systems. In the case of LLMs, for instance, staff could feed confidential data to unaudited models.
- Unintended leakage of confidential data can also arise as a consequence of the LLM training or fine-tuning stage, whereby it is possible that confidential data is recorded by the system and then revealed to users without the appropriate access rights.
- If the model is operated or supported with systems **outside of the environment** where data should be siloed, there might be greater risk of weak links in the process. For example, data leak might occur if data is sent to an externally hosted server to fine-tune a model.
- The use of **code generated by LLMs**, if not properly audited, can also introduce new security risks.

Possible Mitigation Approaches

- Practitioners in the finance industry already have access to a **set of software security and integrity principles**. Employing these principles can mitigate many security issues of LLM applications. In this sense, appropriate **training** remains an important layer of defense.
- To increase robustness against attacks, practitioners can **limit LLMs to specific activities**. In virtual assistant systems, safeguarding mechanisms such as **limited prompts** can be employed to limit the possibility of sensitive information about the company being leaked. However, some noted that this approach has proven difficult to implement.

Fairness and Bias

AI bias refers to the skewed or unrepresentative nature of the data used to train an AI model, or to systemic errors in machine learning algorithms that produce unfair or discriminatory outcomes. Awareness of bias in AI systems that might induce unfair outcomes is clearly important. This involves evaluating how biases in data and algorithms interact with human decision-making processes throughout the AI lifecycle. Fairness, in particular, is a long-term interest and research focus in insurance and banking as the outcomes of services in these sectors can affect differentially members of distinct socioeconomic groups.

Practical Challenges

- Data and algorithmic biases can affect AI models' behaviour in a way that is considered unintended or unfair. A biased output can impact both consumers (for example, via decisions on creditworthiness and insurance pricing) and organisations (for example, businesses using biased predictive analytics might misinterpret market trends, resulting in poor investment decisions or the misallocation of resources) if the source of the bias has not been fully vetted and understood. However, there is not always an objective way to identify a bias, as the definition of a biased system can vary depending on the perspective and circumstances.
- AI model bias is not necessarily due to a flawed model architecture. Bias in an LLM can derive from pre-existing patterns in the data that reflect, for instance, societal and human biases, or that carry insufficient information on under-represented groups, preventing the model from learning effectively. Although this type of bias can merely reflect the nature and shortcomings of available data, its embedding in an LLM has the potential to crystallise them and amplify their impact, making it systematic and large-scale.
- **Agreeing on and implementing fairness principles** is not trivial for simple machine learning models, and it is even more complex for black-box systems such as LLMs.
- Financial data from **different geographical areas** can be characterised by distinct patterns, trends and regime shifts. This can make it difficult to develop a general-purpose foundational model embedding a fair global representation.
- Mitigating biases in financial data is not straightforward, as the effects of sample bias and data gaps on model behaviour could be unknowable or subtle. For instance, a model that

analyses job applications to identify candidates that are qualified or more likely to succeed may be trained on past application and career data. If the training dataset lacks data on candidates who were capable but did not apply due to external barriers (e.g. women from certain socioeconomic classes), once these barriers are removed and those candidates enter the applicants pool, the model could still penalise them in favour of applicants similar to past successful ones.

Possible Mitigation Approaches

- Organisations can mitigate data biases by understanding their different types and how they occur throughout the AI lifecycle. By assessing the entire implementation scope, they can minimise potential for accumulated bias to build up into systematic risk. For instance, control mechanisms on a robo-advisor should be primed to spot not just inaccuracy but if the model consistently overlooked a certain area. Properly designed AI systems can actually help reduce human bias and safeguard fairness by rendering decision-making more objective and factual.
- Determining the **right fairness notion with the right fairness metrics** to measure the impact of representation (e.g., demographic and geographic) on different sensitive groups is crucial. For instance, a robo-advisor might exhibit bias towards certain characteristics due to overrepresentation or gaps in training data and based on specific information availability on the person that is being advised.
- Definitions of fairness tend to be **based on principles rather than strictly rules-based**. For example, the European Commission's AI HLEG[17] describes diversity, non-discrimination and fairness in AI as systems developed and used in a way that includes diverse actors and promotes equal access, gender equality and cultural diversity, while avoiding discriminatory impacts and unfair biases that are prohibited by Union or national law.[18] The translation of these principles into operational rules by model developers and users might be subjective and dependent on the specific application, sector and actors involved.[19]

Accountability and Explainability

Accountability is a key concept in AI governance and consists of ensuring that **clear lines of responsibility** exist for AI systems' decisions and outcomes. Transparency principles often require providing stakeholders with **understandable information about AI systems**, including their purpose, their interplay with decision-making processes, and their potential impact.

The concept of accountability is related to, but distinct from, that of explainability of an AI system, which regards making the inner workings of the AI system understandable, shedding light on how it arrives at its outputs. More broadly, the concept of interpretability centres on comprehending the meaning and implications of an AI system's outputs in specific contexts. In the case of LLMs, it is particularly challenging to establish meaningful parameters for explainability and interpretability.[20] Hence, defining the right level of granularity and implementing human-in-the-loop approaches accordingly gain importance.

---

[17] See https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai

[18] See Ethics guidelines for trustworthy AI | Shaping Europe's digital future.

[19] In line with this, AI researchers and practitioners have proposed different methods and benchmarks for the operationalisation of the non-discrimination and fairness principles. In their comprehensive framework for LLM evaluation, Guldimann et al. (2024) consider two widely adopted fairness benchmarks to evaluate an LLM regarding its non-discriminatory behaviour: DecodingTrust, which measures the dependence of the model's judgement over people's income and their sex, and FaiRLLM, which measures the agreement between recommendations made by the model to people of different protected characteristics. They also evaluate the tendency of an LLM to produce biased outputs on three popular bias benchmarks from the literature: RedditBias, differentially evaluating the representation bias of the model with respect to sensitive groups; BBQ, which evaluates the model's tendency for prejudiced answers in ambiguous contexts; and BOLD, consisting of prefixes from Wikipedia articles on potentially sensitive topics, which are then completed by the model and analysed on toxicity, sentiment, and gender polarity. Evaluating 12 state-of-the-art LLMs against a set of benchmarks, Guldimann et al. (2024) find that the models perform especially poorly on the benchmarks concerning fairness.

[20] Miller (2019) argues that what humans typically regard as a "good" explanation is selective (it should explain elements that are the most important or abnormal), contrastive (it should explain why a particular event happened instead of an alternative) and contextual (it should be presented relative to the explainees' beliefs).

Practical Challenges

- A fundamental question is what is a **satisfactory explanation** that an LLM can provide. Explainability can be considered in terms of various factors, such as model weights or source references. **Full explainability**, including a deep understanding of technical parameters such as model weights, remains difficult to achieve given the black-box nature of LLMs and **is not necessarily useful or relevant** for end-users.
- Due to the challenges in obtaining comprehensive explainability, **enhancing accountability** may be more pragmatic and useful. While transparency rules exist, they may not always fully address the complexity and demands for explainability, making enhanced accountability a desirable solution.
- Many major LLM developers currently do not provide **thorough details about their models**. Information about how training data are managed (e.g. in terms of copyright, licenses, and personally identifiable information), how effective companies' guardrails are (mitigation evaluations), and the downstream impact of foundation models (how people use models and how many people use them in specific regions) remain quite opaque.[21]

Possible Mitigation Approaches

- **Retrieval-Augmented Generation (RAG) systems** enhance explainability by infusing the model with trusted data to have it generate more accurate and relevant responses.
- **Chain of Thought Prompting** uses LLMs as reasoning engines, providing explicit reasoning steps that lead to the final answer. By outlining these reasoning steps, the AI system becomes more interpretable.
- **Reasoning and Acting (ReAct)** methods involve LLMs generating both reasoning traces and task-specific actions. This approach helps in understanding the reasoning steps and provides detailed explanations of how the final result was derived.
- **Post-Hoc Explanations** involve feature attribution, where words or tokens are treated as features, and the contribution of each feature to the model's output is assessed. Techniques such as gradient-based methods are used to provide these explanations. However, these methods demand significant computational resources and may not always deliver the type of explanations desired.
- Less technical approaches to explainability include systems that **always point to the source of information used by the AI**. Providing a list of news sources or references can improve transparency but may not fully meet the needs for a satisfactory level of explainability.
- **Watermarking** can increase the visibility and traceability of LLM generated content. This can help prevent recycling LLM-produced training data.
- More fundamentally, institutions should promote a **clear accountability framework**, with a combination of human-in-the-loop and testing. This would positively impact other trustworthiness characteristics such as robustness and security.

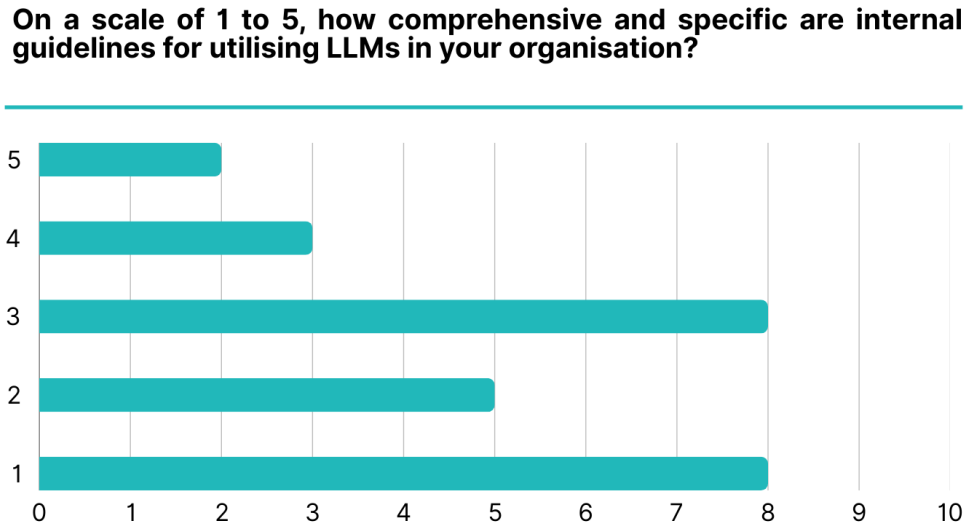
Training and Guidelines

The final topic addressed by the workshop discussion was internal training and guidelines. First, participants were asked to express their opinion on the suitability of any internal guidelines for the use of LLMs currently developed by their organisations.

[21] See Stanford transparency index: https://crfm.stanford.edu/fmti/May-2024/index.html

Workshop Question 5:

On a scale of 1 to 5, how comprehensive and specific are internal guidelines for utilising LLMs in your organisation?



With an average of 2.5 on a scale of 1 to 5, the respondents indicated that not many organisations had developed comprehensive and specific guidelines for the use of LLMs. Following up on this, participants were asked to provide details on the missing elements in their organisations' internal policies.

Workshop Question 6:

What is missing in your organisation's guidelines for the use of LLMs?

The free-text answers can be grouped into three categories: (1) a risk-aware training approach; (2) technical guidance; and (3) clarity on the acceptable use policies and recipes.

1. **Risk-aware training approach includes:**
   - Training to consistently challenge LLM outputs, especially considering risk awareness.
   - Developing thorough evaluation and monitoring of LLM-based systems behaviours (datasets, metrics).
   - Transparency around 'hallucinations' and their likelihood.
   - Guidelines for non-expert users on safeguards measures when using LLMs.
   - Practical privacy implications.
2. **Technical guidance includes:**
   - Guidelines tailored for different use cases and different perspectives (e.g. more technical from an IT charter point of view, structured from a responsibility point of view for business users).
   - Detailed and specific guidelines about confidentiality level of data injected in third-party LLMs.
   - Explainability of the current models.

3. **Clarity on the acceptable use policies and recipes includes:**
    - When to use general customer-facing LLMs or LLMs hidden behind an interface with specific engineering.
    - Development-oriented or programmatic use of LLMs.
    - Dealing with newly developed unknown applications of LLMs.
    - Warnings over capabilities on numeric applications.

A general concern was that guidelines were too broad and did not envisage all potential uses of LLMs. Current use cases are generally limited, so guidelines were not completed as long as the organisations were not able to deploy LLMs for all the use cases. Guidelines often focused on data governance, cybersecurity and privacy issues, but there was little "guidance" on practical examples of tasks facilitated by LLMs.

Some participants pointed out that some best practices referred to consulting management – yet, management was rarely trained in the use of LLMs. Training programs and education could facilitate the practical application of the guidelines.

Practical Challenges

- Many workshop participants mentioned that their organisation initiated training, but pointed out that that was not as intensive as it would be necessary to meet the level of technicality involved in AI oversight and governance. [22] A lot of **guidance is broad and general** due to a lack of clarity on the concrete use cases in the financial sector.
- **Attracting and retaining talent** is considered difficult because LLMs and AI specialists are typically hired by technology firms and startups rather than by the financial industry.

Possible Mitigation Approaches

- Firms and regulators can **train their employees** on appropriate ways to use LLMs, in specific use cases, establishing approaches to check outputs and designing other methods to reduce errors.
- **Banning internal use** because of security or privacy concerns cannot necessarily be ruled out, but can lead to undesired results since it might induce employees to use LLMs on their personal computer, potentially exacerbating the risk of data leaks.
- Some financial organisations have a **team dedicated to the audit of models** which could support in training activities.
- Some believe that, in the future, successful integration of LLMs enabled by appropriate training can create a "**symbiotic" relationship between human intelligence and machine capabilities**, leading to decisive productivity gains.

Carbon Footprint of LLMs

LLMs require a significant amount of energy for their training and inference, which poses questions around their sustainability. As organisations devote greater attention to their carbon footprint to project responsibility to clients and society, understanding the energy consumption of LLMs becomes increasingly relevant not only for their developers but also for their financial sector adopters.

Patterson et al. (2024) reported that training GPT-3 with 175 billion parameters consumed 1287 MWh of electricity. This represents less than 1/100,000 of the global data centre electricity consumption in 2022, which in turn accounted for around 1% of global final electricity demand.[23] Considering that the training of an LLM is a one-off occurrence and the number of LLMs is not

[22] See also "The impact of AI on the workplace: Main findings from the OECD AI surveys of employers and workers", available at: https://www.oecd.org/en/publications/2023/03/the-impact-of-ai-on-the-workplace-main-findings-from-the-oecd-ai-surveys-of-employers-and-workers_ad686e91.html

[23] See Data centres & networks - IEA.

expected to scale up significantly, these figures still appear negligible in light of the volume of work that LLMs are expected to accomplish if adopted globally.

Once models are deployed, inference – i.e. the model's production of output in response to the queries submitted by users – may consume a considerable amount of energy. Google reported that 60% of AI-related energy consumption from 2019 to 2021 stemmed from inference – not dissimilar to the proportion ascribed to training – but this ratio might increase depending on the future trajectory of LLM adoption. With ChatGPT averaging already 123.5 million daily active users and processing over 1 billion queries per day, the technology might become a more significant source of electricity demand.

Nevertheless, it is also important to note that considering the energy consumption from the use of LLMs – and, more generally, AI – in isolation might be misleading. A sound cost-benefit analysis would need to account for the energy savings enabled by the use of AI-based tools as they replace legacy, less efficient technologies in fulfilling certain tasks.

Practical Challenges

- As the use of LLMs spreads within organisations, employees will be eager to use LLMs for more applications. A single request in ChatGPT consumes around 10 times more energy than a Google search (see de Vries, 2023), but estimates again vary greatly.[24]
- Actual and projected energy consumption generated by the use of LLMs is **difficult to estimate** – there is considerable uncertainty around these estimations.

Possible Mitigation Approaches

- Cost-benefit analyses on the use of LLMs could include **carbon footprint metrics**. An **estimation of the volume of LLMs use –** and when possible a quantification of the corresponding **energy consumption** – could be done within each organisation.
- The carbon footprint of LLMs can vary significantly depending on the model size and in particular the number of parameters used for inference (the difference in energy consumption between large and smaller, more efficient LLMs can be of a factor of 100). Research effort is done to reduce the size of these models (see, for example, the "BabyLLM" challenge and claims around Chinese start-up DeepSeek's models consuming less energy).[25] Choosing a more **parsimonious model** that has been fine-tuned for specific tasks can have a large impact on carbon footprint.

[24] See https://www.thedigitalspeaker.com/greener-future-importance-sustainable-ai/.
[25] See https://babylm.github.io/ and https://www.forbes.com/sites/wesleyhill/2025/02/03/chinas-deepseek-ai-reshapes-global-energy/.

5. Discussion

5. Discussion

The use cases for LLMs are evolving but practical challenges remain

The Alan Turing Institute's FAIR project released in March 2024 a report on "*The Impact of Large Language Models in Finance: Towards Trustworthy Adoption*" to capture collective insights from high-street banks, regulators, and other financial services industry stakeholders.

One year may be brief for most industries, but it is transformative in finance and AI, especially AI in finance. When we released last year's report, we were exploring the early promises and challenges of LLMs. Now, we started seeing the evolution of these technologies, adopted by different organisations, transforming both the development and governance of products.

It is important to note that LLMs, compared to other machine learning tools adopted by the finance industry, **are still young and changing fast**. It takes time to analyse their capabilities and assess their long-term impact. And, in some cases, LLMs are not the right tool to be used. Both reports revealed that the complexity of LLMs demands a **delicate balance between innovation and compliance**. Financial institutions are increasingly prioritising the definition of accountability while ensuring decisions derived from LLMs are intuitively or directly explainable and can withstand scrutiny from regulators and stakeholders.

The evolution of LLMs has highlighted the importance of **robustness and resilience** in financial applications. We observe an increasing need for **robust benchmarks and risk assessment methodologies that can handle unpredictability**. Models must function reliably under diverse conditions, including **handling adversarial inputs and maintaining performance throughout data shifts**. This has led to an increased focus on **stress-testing AI systems**, akin to what is already standard practice for traditional financial risk management tools.

Privacy has also emerged as a critical focus area. As LLMs handle sensitive financial data, organisations are adopting advanced techniques like **federated learning and differential privacy to protect customer information**. Ensuring privacy without compromising the utility of AI tools remains an ongoing challenge.

Also, the **ethical implications** of LLM adoption continue to evolve. Issues such as mitigating bias in financial decision-making and ensuring equitable outcomes for all stakeholders in financial services are receiving heightened attention.

Research is moving fast

The adoption of LLMs in finance has not only changed the practices of industry professionals, but also the direction of academic research. The current academic literature in finance is moving towards a **broader use of LLMs**, as can be seen in the figure below, which counts the number of research articles in finance that use LLMs. LLMs offer the ability to analyse and integrate large amounts of new unstructured data and to easily simulate the behaviour of economic agents. The study of the economic and financial consequences of LLM adoption is also a research topic in itself.

Figure: Number of research papers released quarterly on LLMs and Finance. Source: FaIR, Institut Louis Bachelier, based on SSRN & ARXIV publications.

Initially focused on basic **natural language processing** (NLP) applications such as sentiment analysis and financial document classification, research using LLMs in finance has evolved with recent advances in LLMs, which can now handle much **more complex tasks** (Eisfeldt & Schubert, 2024). Specialised models such as FinBERT and BloombergGPT have proven particularly effective for finance. In particular, the introduction of ChatGPT has democratised the automation of text analysis tasks. The use of retrieval-assisted generation (RAG) systems, which allow verifiable data to be incorporated into language models and combined with financial databases, now improves the quality of model output.

Recent research in finance reflects a growing interest in advanced AI applications in areas such as **risk management** – e.g. for fraud detection (Jiang, 2024) or **credit risk analysis** (Sanz-Guerrero and Arroyo,2024) – **customer communication** (Lo and Ross, 2024; Fedyk et al., 2024) and **investment** (Chen et al., 2022 ; Ko and Lee, 2024; Jha et al., 2024), including sustainable finance (Brière et al., 2024 ; Spacey et al., 2024). **Interpretability** is a priority in recent research. Chang et al. (2023) and Touvron et al. (2023) show the importance of model transparency for wider adoption in regulated environments. Current research into LLM applications also demonstrate a growing interest for **model robustness and scalability** (Chen et al., 2024; Zhao et al., 2024), bias and hallucination reduction (Wei et al., 2024). However, challenges remain regarding **data governance, confidentiality, and privacy** (Yip and Balagué, 2023). The rise of federated learning should help preserve the security of sensitive information while maximising model efficiency (Aldasoro et al., 2024; Wen et al., 2023). A reflection is in progress concerning the use of more elaborate and multi-dimensional benchmarks for LLMs evaluation (Liang et al., 2023). Right now, benchmarks mostly test against a single goal – accuracy. But more holistic benchmarks may help us understand the trade-offs between the various performance metrics of LLMs, for example accuracy, bias or toxicity.

Regulatory framework

In the EU, the **AI Act**[26] sets out requirements and obligations for AI developers and deployers across sectors for specific uses of AI. The Act follows a **risk-based approach**. It prohibits AI practices posing "unacceptable" risk, such as social scoring by governments, and determines a list of "high-risk" applications such as credit decisioning. For the latter, the AI systems must adhere to specific requirements and their deployers and providers are subject to certain obligations. Foundation models, including LLMs, are captured under the notion of **general-purpose AI models (GPAI), further split into GPAI models likely to raise systemic risk or not**. The Act introduces **transparency requirements** for GPAI models and additional **risk management obligations for high-impact**, highly capable GPAI models that might pose systemic risk. These additional obligations include self-assessment and mitigation of systemic risks, reporting of serious incidents, conducting test and model evaluations, as well as cybersecurity requirements. The GenAI tools

[26] Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence.

based on LLMs will have to comply with these requirements and EU copyright law, including disclosing that the content was generated by AI, designing the model to prevent it from generating illegal content, and publishing summaries of copyrighted data used for training. [27]

The regulatory requirements set out in the Act for GPAI models and AI systems will be further specified through technical requirements and standards. As part of this effort, the newly established EU AI Office is collaborating with stakeholders to draw up a **general-purpose AI Code of Practice**, expected to be published in May 2025. The Code of Practice aims to facilitate the proper application of the AI Act's rules for GPAI models, including **transparency and copyright-related rules for model providers**. For providers of GPAI models with systemic risks, the Code should also detail a taxonomy of systemic risks, risk assessment measures, as well as technical and governance mitigation measures. [28]

It is also important to recall that beyond any specific AI rules, the use of LLMs and other forms of AI in the financial sector is also disciplined by **pre-existing legislation**, which applies irrespective of the technology used. In the EU, for instance, MiFID II includes requirements for investment firms and trading venues engaged in algorithmic trading and high-frequency trading, activities that can also leverage AI systems. Relatedly, in May 2024, ESMA issued a statement to provide initial guidance to firms using AI when they provide investment services to retail clients, including e.g. via LLM-powered chatbots and virtual assistants, so that they can ensure compliance with the MiFID II requirements.[29]

Looking at global regulatory approaches, the vast majority of the 49 jurisdictions recently surveyed by the Organization for Economic Co-operation and Development (OECD) reported that they had appropriate regulation in place, while acknowledging potential gaps and the need for more general guidance. Additionally, almost all surveyed jurisdictions have introduced some form of policy that covers AI in (parts of) finance.[30]

From a wider **financial stability perspective**, a greater usage of pre-trained models, together with an increasing importance of specialised hardware and cloud services for AI development, creates more **third-party dependencies**. In its recent report on the financial stability implications of AI, the Financial Stability Board highlighted that the greater reliance on and market concentration among LLM providers could increase systemic third-party dependencies in the financial sector, which, going forward, might become a concern for regulators from an operational vulnerability perspective.[31]

Overall, the use of AI and LLMs in finance is constantly evolving and appears increasingly relevant, which makes engagement with external stakeholders – including through workshops like the one distilled in this report – particularly relevant for regulators.

[27] See AI Act | Shaping Europe's digital future (accessed on 29 November 2024).
[28] See European Commission (2024), First Draft of the General-Purpose AI Code of Practice published, written by independent experts (November).
[29] See ESMA (2024).
[30] See OECD (2024).
[31] See Financial Stability Board (2024).

6. References

6. References

Aldasoro, I., Gambacorta, L., Korinek, A., Shreeti, V., & Stein, M. (2024). Intelligent Financial System: How AI is Transforming Finance. Technical report, Bank for International Settlements.

Brière, M., Keip M., Le Berthe T. & Nuriyev M. (2024). Artificial Intelligence for sustainable finance: why it may help. Available at SSRN 4252329.

Chang, Y., Wang, X., Wang, J., Wu, Y., Zhu, K., Chen, H., Yang, L., Yi, X., Wang, C., Wang, Y., et al. (2023). A survey on evaluation of large language models. arXiv preprint arXiv:2307.03109.

Chen, Y., Kelly, B.T. and Xiu, D., 2022. Expected returns and large language models. *Available at SSRN 4416687*.

Chen, Y., Zhang, Z., Han, X., Xiao, C., Liu, Z., Chen, C., Li, K., Yang, T., & Sun, M. (2024). Robust and Scalable Model Editing for Large Language Models. arXiv preprint arXiv:2403.17431. https://arxiv.org/abs/2403.17431.

De Vries, A. (2023). The growing energy footprint of artificial intelligence. Joule, Volume 7, Issue 10, Pages 2191-2194, https://doi.org/10.1016/j.joule.2023.09.004.

Eisfeldt, A. L., & Schubert, G. (2024). AI and Finance. NBER Working Paper Series. National Bureau of Economic Research.

ESMA (2024). Public Statement on AI and investment services. https://www.esma.europa.eu/sites/default/files/2024-05/ESMA35-335435667-5924__Public_Statement_on_AI_and_investment_services.pdf

ESMA (2025). Warning on the use of AI for investing. https://www.esma.europa.eu/sites/default/files/2025-03/ESMA_Warning_on_the_use_of_AI_-_EN.pdf

Fedyk, A., Kakhbod, A., Li, P., & Malmendier, U. (2024). ChatGPT and Perception Biases in Investments: An Experimental Study. Available at SSRN 4787249.

Financial Stability Board (2024). The Financial Stability Implications of Artificial Intelligence. https://www.fsb.org/2024/11/the-financial-stability-implications-of-artificial-intelligence/

Guldimann, P., Spiridonov, A., Staab, R., Jovanović, N., Vero, M., Vechev, V., Gueorguieva, A., Balunović, M., Konstantinov, N., Bielik, P., Tsankov, P., & Vechev, M., "COMPL-AI Framework: A Technical Interpretation and LLM Benchmarking Suite for the EU Artificial Intelligence Act". arXiv preprint arXiv:2410.07959. https://arxiv.org/abs/2410.07959.

Jha, M., Qian, J., Weber, M., & Yang, B. (2024). ChatGPT and Corporate Policies. Technical report, National Bureau of Economic Research.

Jiang, L. (2024). Detecting scams using large language models. arXiv preprint arXiv:2402.03147.

Ko, H., & Lee, J. (2024). Can ChatGPT improve investment decisions? From a portfolio management perspective. Finance Research Letters, 64, 105433.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A. and Newman, B., 2023. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Lo, A. W., & Ross, J. (2024). Can ChatGPT Plan Your Retirement?: Generative AI and Financial Advice. Available at SSRN 4722780.

Maple, C., Sabuncuoglu, A. (2024). The impact of large language models in finance: towards trustworthy adoption. The Alan Turing Institute.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, Volume 267, Pages 1-38.

Novelli, C., Taddeo, M. & Floridi, L. (2024) Accountability in artificial intelligence: what it is and how it works. *AI & Soc* **39**, 1871–1882.

OECD (2024). Regulatory approaches to Artificial Intelligence in finance. https://www.oecd.org/en/publications/regulatory-approaches-to-artificial-intelligence-in-finance_f1498c02-en.html

Patterson, D. and Gonzalez, J. and Le, Q. and Liang, C. and Munguia, L. and Rothchild, D. and So, D. and Texier, M. and Dean, J. (2024). Carbon Emissions and Large Neural Network Training. arXiv preprint arXiv: 2104.10350.

Sanz-Guerrero, M., & Arroyo, J. (2024). Credit Risk Meets Large Language Models: Building a Risk Indicator from Loan Descriptions in P2P Lending. arXiv preprint arXiv:2401.16458.

Spacey Martín, R., Ranger, N., Schimanski, T., & Leippold, M. (2024). Harnessing AI to assess corporate adaptation plans on alignment with climate adaptation and resilience goals. *Available at SSRN 4878341*.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Roziere, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Wei, J., Yao, Y., Ton, J-F., Guo, H., Estornell, A., & Liu, Y. (2024). Measuring and Reducing LLM Hallucination without Gold-Standard Answers. arXiv preprint arXiv:2402.10412. https://arxiv.org/abs/2402.10412.

Wen, J., Zhang, Z., Lan, Y., Cui, Z., Cai, J. and Zhang, W. (2023). A survey on federated learning: challenges and applications. International Journal of Machine Learning and Cybernetics, 14(2), pp.513-535.

Yip K., Balagué C. (2023). ChatGPT: research evidence-based controversies, regulations and solutions. Good In Tech Rapport, Equipe de recherche Good in Tech; Institut Mines Télécom Business School; Sciences Po; Fondation du risque, Institut Louis Bachelier. 2023, pp.1-27. hal-04705791.

Zhao, Y., Yan, L., Sun, W., Xing, G., Wang, S., Meng, C., Cheng, Z., Ren, Z., & Yin, D. (2024a). Improving the Robustness of Large Language Models via Consistency Alignment. arXiv preprint arXiv:2403.14221. https://arxiv.org/abs/2403.14221.