Auditing the auditor: evaluating and implementing LLMs and agentic AI performance within heterogeneous data, zero fault-tolerant environment of financial audit and analysis

Jingkun (Charly) Zhu
King's College London / Tsinghua University
Fast Audit AI / FirmView AI
London, UK
charly.zhu@firmview.ai

Wojtek Buczyński
University of Cambridge / London Business School
Fast Audit AI / FirmView AI
London, UK
wojtek.buczynski@firmview.ai

ABSTRACT

Financial services are highly invested in being a cutting-edge, tech-driven industry and have the resources to experiment with emerging AI technologies – including LLMs and, more recently, agentic AI.

The main challenge is of course LLM confabulations, commonly known as hallucinations. They are completely incompatible with the industry's focus on accuracy and correctness. For this reason alone, human oversight is essential – another, even more fundamental reason is regulation.

This is how we arrive at the LLM and agentic AI nexus in the financial services: considerable promise and non-negotiable requirement for human oversight, governance, and audit.

In this paper we focus on two agentic workflows: (i) semi-automated financial data collection from trusted sources and (ii) semi-automated investment research generation.

Our contributions to the workshop are:

- A discussion on how to evaluate and audit the work of LLMs and agentic AI from a joint academic-pragmatic perspective using two real-world, in-depth use cases.
- A versatile, comprehensive and human-centred audit and evaluation framework for LLMs and agentic AI.

CCS CONCEPTS

Human-centered computing → Human computer interaction (HCI) → HCI design and evaluation methods

Human-centered computing → Human computer interaction (HCI) → Interactive systems and tools

KEYWORDS

Artificial Intelligence (AI), agentic AI, Large Language Models (LLMs), HCI, financial services, audit

1 Introduction

Artificial Intelligence (AI) is widely considered – both in academia [4, 10, 15, 18] as well as the financial services industry [3, 7, 13, 19] as a technology with the potential to disrupt the industry. It may disrupt the long-established hierarchy of financial services firms by enabling technology-driven, agile new players (FinTechs) – including two AI FinTech startups we launched ourselves – to enter the market and compete with the incumbents[2, 6, 20, 23].

The public launch of ChatGPT in November 2022 was a seminal event. The world of professional services took note almost immediately. In June 2023 McKinsey released a report titled "The economic potential of generative AI" [16] which contains estimates of gen AI's economic impact that became a reference point in many professional industries: *"Generative AI's impact on productivity could add trillions of dollars in value to the global economy."*.

Finance-related use cases followed almost immediately: risk and compliance [21] ,M&A [5], sales [1] and ultimately wealth and asset management (WAM) [17]The problem is that while adoptions rates are reportedly very high (and look very good in industry surveys [25] or in annual reports), what we're hearing is that Large Language Models (LLMs) are frequently used to create low-quality output referred to as "workslop" [14]or to simplify routine tasks such as summarising documents or rephrasing e-mails.

Agentic AI could turn out to be as transformative as LLMs and working in tandem they could develop value-adding, beneficial emerging properties that neither could deliver on its own. However, it is yet to achieve the level of market penetration of generative AI – possibly because gen AI has captured global imagination to such extent, that there is little "bandwidth" left for another cutting-edge technology. A more technical and business explanation could be that gen AI is highly visible and direct, while agentic AI is more abstract and technical; the former is about instant results while the latter is more about orchestration. For the time being we believe it is fair to consider agentic AI technology to be more niche and focused on more technical users; its general public breakthrough is yet to come.

We consider AI implementation – including deployment, everyday functioning (BAU) and oversight – an inherently interdisciplinary endeavour, with thorough understanding of stakeholders, their needs, goals and constraints as the starting point.

Our approach is based on the three pillars: use case → stakeholders → technology.

With the two use cases combining years of academic research with a practical, real-life approach (as both use cases are taken from AI FinTech startups we launched) we want to discuss and examine how to combine constantly increasing capabilities of LLMs and agentic AI with human oversight and audit for reliable, value-adding outcomes.

2 Human evaluation and audit of LLMs and agentic AI

2.1 Financial audit

Financial audit is a unique industry in many ways. It is not regulated like other financial services and not even considered to be a part of financial services proper, yet its efficient functioning is crucial to any developed economy. Audit has a unique distinction of a "public trust industry". It could be argued that auditors are as essential to successful functioning of developed economies as retail banks or credit card companies – although less visible.

Audit, at its core, is not exceptionally challenging intellectually. It is, however, exceptionally laborious, increasingly complex – and, importantly, it has not changed much in decades. While screens and spreadsheets have replaced paper financial statements, audit remains largely about looking at numbers, analysing them, and reconciling them. Over the years, as the size and complexity of business has been growing, so has complexity of the audits – as well as their costs, which have been rising at astonishing rates [12].

The entire premise of our audit startup is that parts of the audit workflow appear to be promising candidates to be augmented and radically enhanced by AI; up to 95% on some tasks according to our research. We focused on a specific task within the audit planning phase called anomaly detection, or in other words: finding numbers that do not fit.

In financial statements – just about any corporate financial statement, regardless of the industry, region or size of the entity – an anomaly is going to be one of three things:

- Something that is perfectly legitimate and explainable in the context of company's activities (for example an acquisition of another entity – those frequently generate some financial statement anomalies in the early years);
- A perfectly legitimate company experiencing financial turmoil such situations are often indicated by abnormal numbers;
- An indication of financial irregularities of illegitimate nature: "creative accounting" etc.

The challenge is that identifying an outlier – something Machine Learning does very well – is of limited value to the auditors; without context, without explanation, without knowing "why is this an outlier?" there is little they can do with this insight. This is exactly where LLMs can add substantial value – they can provide context and / or offer some plausible, factual explanations as to why a given number is what it is – and particularly, whether the cause may be perfectly legitimate; or whether this is something that needs to be investigated thoroughly as potential wrongdoing.

Unfortunately, in the current state of generative AI, there is no way to guarantee full factual accuracy in RAG (retrieval augmented generation) tasks. This creates a seemingly irreconcilable friction between the stakeholders: the auditors, the audit oversight bodies (Financial Reporting Council in the UK) and us, the AI vendors. At the same time audit is hitting a certain wall as an industry: the audits are becoming only ever more complex and their cost is becoming increasingly burdensome, especially for small and mid-sized businesses (SMEs).

Can this issue be resolved somehow? We developed a couple of ideas; importantly, they are not purely technological – the fundamental part of our approach is human-centricity, auditability and transparency.

Firstly, we started by strategically selecting the LLM itself. We went for an open-source model to avoid dependencies on a third party (cost, opaqueness, behind-the-scenes model updates etc.). As the initial release of our system was targeted at British and European clients, we chose a European LLM for a better contextual fit. In the future, as we expand coverage, we are considering employing a US LLM for the American clients and a Chinese LLM for the Chinese ones.

Secondly, we extensively trained the model in-house – to the extent that we had the complete LLM running on our local machines. We fed the LLM with the curriculum of professional accountancy qualifications and then fed it thousands of financial statements, earnings call transcripts and other relevant inputs from the universe of the companies we cover.

We then combined the LLM with an agentic orchestration layer: a task-level agent decomposes the user request, dispatches specialised nodes/tools (retrieval, reading/extraction, reconciliation checks), and decides the next step based on structured intermediate outputs.

The system is coordinated by a Planner agent that produces an explicit, executable next-step decision, following the broader "plan–execute" and "planner–executor" separation that has been shown to improve controllability and efficiency in tool-augmented LLM systems.

Upon receiving a user query, the Planner can invoke evidence-acquisition agents, reflecting a modular "router-to-experts" design common in tool-augmented systems: (i) Internal Data Agent (queries internal APIs for structured financial records and analysis outputs), (ii) Web Agent (targeted external search with ranking and normalization), (iii) Reader Agent (long-document reading and extraction with provenance), and (iv) Analysis Agents (domain checks, reconciliations, and consistency tests). This modular tool-usage view aligns with prior work showing the value of composing LLMs with external tools and expert modules. Each invocation yields one or more Atomic Evidence Packs (AEPs). An AEP is the minimal auditable unit produced by any tool call or agent operation. Each pack represents either (a) a single piece of externally observed evidence (e.g., a retrieved passage, a database row, a web page snippet, a computation result), or (b) a single derived analytic artifact whose inputs are

explicit AEP references. This "atomicity" mirrors the broader trend toward decomposing generation into verifiable units (e.g., citation evaluation benchmarks and quote-level evidence) [27] and toward evidence interfaces in auditable agents[32]. AEP creation is mandatory for any tool result used downstream, converting raw tool outputs into provenance-rich, inspectable units and avoiding "citation-by-assertion" where references exist but the claim is not actually supported.

Each Atomic Evidence Pack (AEP) stores: (1) content: the extracted fact/snippet/table/value, or a derived analytic conclusion whose inputs are explicit AEP references; (2) journey: a logged sequence of model/tool steps that produced the AEP (for post-hoc debugging); (3) provenance: source type (internal API, web page, document), source identifier, retrieval parameters/query, timestamp, and tool/agent version; (4) scope metadata: entity, time period, document section, and audit-relevant tags; (5) stance edges: optional typed links to other AEPs with 'supports'/'contradicts'; and (6) justification: a short rationale (with quote-level pointers when feasible) for stance links.

When the Planner judges that evidence coverage is adequate, it compiles an evidence bundle (a) subset of AEPs plus key support/contradiction links) and hands it to a Writer agent to draft the user-facing response. This explicitly separates (a) evidence acquisition and (b) narrative synthesis, which mirrors "draft–verify/revise" families of methods that reduce unsupported content by inserting a revision loop rather than treating the first draft as final. Importantly, the Writer is permitted to reject the bundle: if evidence is insufficient, internally contradictory without an explicit resolution path, or not authoritative for the required claim strength, the Writer returns a structured critique to the Planner indicating missing evidence types (e.g., "need authoritative source," "need numeric confirmation," "need time-period disambiguation"). This operationalizes critique as an interface between modules and is consistent with retrieval-and-critique paradigms that treat critique as a first-class control signal.

The Planner terminates evidence acquisition when one or more of the following holds. First, evidence sufficiency: a complete answer can be produced with AEP citations covering all key claims and with acceptable source supportiveness. Second, evidence saturation: additional retrieval produces redundant evidence or fails to increase claim coverage, consistent with modular/looping RAG flows where retrieval is iterated only when it adds marginal value. Third, bounded search: resource constraints (tool budget, latency, or policy limits) are reached; the system then produces a qualified answer (with explicit uncertainty and evidence gaps) or escalates to human review depending on the operating policy, reflecting the cost/efficiency motivations emphasized in token- and tool-efficient augmented-LM frameworks.

In high-stakes finance workflows with little to no margin for error, we do not rely on the agent alone to control hallucinations, because small factual errors can compound across steps. We therefore adopt a tiered review policy: (i) for numeric facts and extracted financial data (e.g., values, units/currency, periods), human verification is mandatory; (ii) for templated narrative completion where the outline and evidence are fixed, we use spot checks; and (iii) for consequential analytic claims, we require full human read-through. To reduce reviewer burden, we additionally use a cheaper verifier model to flag inconsistencies for human escalation.

The entire system is "packaged" into a user-facing chatbot we called Charly – "Charly the chatbot" aka CtC – which was added to the existing, Machine Learning-based anomaly detection system. This made the ultimate synergy: quantitative analysis of financial statements augmented by qualitative insights from CtC.

CtC is designed not to give definitive statements (something conventional LLMs are very good at, even if they're confabulating) and to speak in terms of what could be the likely explanation. Furthermore, we tried to make it so that whenever possible CtC presents a range of plausible explanations. This "inconclusivity" is important when auditors are the key stakeholders. Auditors need to retain their intellectual autonomy and professional judgment – fundamental tenets of their industry – with our solution positioned as an analytical support tool.

2.2 Equity analysis

Equity research is a fundamental part of the investment industry – it is also tightly regulated. It serves two main purposes:

It provides a snapshot of the current situation of the covered company including all material information, developments and financial performance, prepared by an expert analyst.

It provides share price forecast for a fixed time horizon (usually 6 – 12 months) along with a simple recommendation: Buy, Sell or Hold.

Typically, a research report is a combination of free-form text and financial data. It is the analyst's personal judgment to decide what information they consider material and relevant. The price forecast is the ultimate expression of the analyst's judgement and should be a logical conclusion of the facts, numbers and opinions included in the report.

We find four practical use cases for agentic LLMs in investment workflows: (1) automated news aggregation on companies and geopolitical context; (2) automated aggregation of user/consumer reviews and commentary; (3) semi-automated financial data collection from trusted sources; and (4) semi-automated investment research generation. In this submission we focus on (3) and (4), because they demand the strongest auditability, provenance, and human review policies.

(3) and (4) are perfect use cases to utilise LLM + agentic AI combo. Information that can be material and have impact on a company's performance, and, by extension, on its share price is vast, unbounded and heterogenous. It could be general economic and market developments (global or local); it could be industry-specific impacts; it could be political or regulatory developments; lastly it could be random, unforeseeable events like Covid. While research reports and financial statements tend to be highly structured and orderly, the starting point is subjective information sourcing, data selection and filtering. Historically, this has always been a domain of human analysts – highly-paid professionals who honed their talents and instincts through years of practice. As the job has a degree of subjectivity and judgement call – how large or

small depending on the investment firm and its research frameworks – it has been considered highly "cerebral" and generally safe from AI disruption. Then came LLMs, followed by agentic AI, and they started to change everything.

Currently just about any LLM can produce a professional-looking report on any entity, of any chosen length, fine-tuned to the user's specification. The problem is that such reports usually just *look* smart, but provide no original insights – while coming with the risk of making up false statements. Finance professionals recognise that such reports have little or no value, and require thorough proof-reading and fact-checking from an expert, which defeats their purpose. However, all the material and meaningful information is online, and it is often possible to arrive at new, original conclusions that may impact the share price – so the tool is not misplaced and the use case does make sense; it just needs very thoughtful fine-tuning.

Two design decisions emerged early through iteration. First, we moved away from a "standalone" LLM approach toward a combined LLM + agentic workflow, which produced immediate improvements in scope and depth by decomposing work into tool-using subtasks. Second, we found that whether the user interacts with a live system (chat-like) versus a static artifact (a report) fundamentally changes the risk profile and control strategy. Static report generation enables explicit verification gates and a final expert review before release, which in turn makes it feasible to reduce unsupported claims below the workflow's release threshold.

Because unsupported claims and small numeric errors can compound across multi-step pipelines, we do not treat "model self-checking", sometimes effective for less critical use cases, as sufficient. Instead, we enforce verification gates: (i) automated supportiveness checks for whether cited evidence actually supports a claim [33] (ii) lightweight cross-checks for high-risk numeric fields; and (iii) mandatory expert review at release time for the static report artifact. This design follows the broader trend of draft–verify loops [34] while recognising that regulated workflows require explicit documentation and logging for post-hoc inspection[24, 31].

When multiple sources disagree (common in finance due to reporting conventions, currencies, and revisions), we represent support/contradiction relations explicitly as typed edges between AEPs, rather than silently resolving conflicts inside the model. Conceptually this aligns with bipolar argumentation structures [8, 9] and recent retrieval-plus-argumentation directions for contestable verification[29]. The practical benefit is stakeholder-facing: compliance can see what evidence was competing, business can see what was chosen and why, and engineers can debug failure points.

We also designed an agentic auditing assistant whose primary technical objective is auditability: the ability to reconstruct, inspect, and contest how a final answer was produced. Concretely, every intermediate step that can influence the final answer—retrieval results, tool outputs, derived computations, and conflict resolution decisions—is captured as an AEP. AEPs are accumulated into an evidence set that is reused across turns in the same conversation; in future work, the set can be upgraded into a vector-indexed evidence store, aligning with the standard RAG pattern of maintaining a non-parametric memory that can be updated without retraining[22].In addition to being a storage object, an AEP is also a control object: subsequent reasoning steps must explicitly (i) cite AEP ids when making claims, and (ii) optionally attach quote-level anchors where the modality permits (text passages, document extracts). This is motivated by evidence-grounded systems that emphasize quote-ability and reproducibility as part of the evidence interface.

All outputs from AI agents for equity analysis are being logged and they are being reviewed – alongside the final, proposed report content – by a human analyst. Given the static nature of equity research reports and the severe downside of allowing any hallucinated inputs we put human in the loop at various stages of the equity report generation – chief among them being the final approval. We see this approach as a win-win: the equity research process is still being enhanced by at least an order of magnitude (possibly two), while we have a human specialist acting as the final quality control "approver"; separately, it also addresses any potential regulatory concerns should we become a regulated entity one day.

3 Evaluation and audit of LLMs and agentic AI in practice

As finance professionals and present-day AI FinTech founders our perspective on LLMs and agentic AI is probably quite different to the way some academics approach it. We have limited interest in theory – our approach is "how can / will this work in practice?"; "how can this add value?"; "what are the compliance and regulatory considerations?". Furthermore, as practitioners, we understand first-hand where things could go wrong and how we could go about avoiding that.

Based on our experiences as industry practitioners, researchers and now AI FinTech entrepreneurs, we propose the following framework. It is a framework we follow at both of our AI FinTech startups:

1. Define and understand the stakeholders.

Stakeholders need to be clearly defined, and it generally shouldn't be one, narrow group of people. Fully understanding stakeholders' needs, goals and constraints is fundamental to a successful deployment of any technology, not just AI.

The considerations applicable to end-user (client) stakeholders are just as important on the deployer (vendor) side. We very strongly believe that successful launch and deployment of an AI system – LLM, agentic, ML or otherwise – requires an interdisciplinary team, which in addition to technologists also includes subject matter experts on the business side, compliance and regulatory experts, data specialists and even ethics experts when applicable.

2. Define optimal workflows.

Designing an optimal workflow is key. We think the "plug and play" approach pushed quite aggressively by LLM chatbot or "copilot" vendors (and consultancies) is naïve, risky and unlikely to yield meaningful results. Unfortunately, to create a value-

adding solution usually a meticulous, iterative, consultative process is required. Our approach is two-pronged: use case (broken into "bite-sized" standalone subtasks) + stakeholders. Optimal technologies can be considered a third prong, but they are always the tool and a mean, never a starting point. In both use cases we presented above the workflows are multi-step and utilise different technologies for different subtasks. We also have human evaluators at multiple stages. Importantly, human oversight is permanent and ongoing by design.

3. **Clearly delineate the use cases where end-users will be interacting with the LLM directly (e.g., CtC) as opposed to those where there has to be a "human filter" (e.g., equity analysis).**

While the equity research workflow might sound similar to the process for financial audit, there is a fundamental difference: equity research reports are static documents with a certain "shelf-life" (usually a quarter or more) while an auditor's discussion with a chatbot is live. That difference meant that despite sounding similar, our LLM + agentic AI orchestrations for financial audit and for equity analysis differed a great deal.

4. **Keep reviewing and fine-tuning the LLM on permanent basis. There is no such thing as a "steady state LLM" as far as we're concerned.**

5. **Build complex AI systems (and all the systems we're discussing here are complex) with radical transparency as a guiding principle.**

Our approach towards transparency has two main pillars: technology and people.

- We set up the LLM + agentic AI combo in such way that agents weren't just searching for information – they also had to justify its importance, validity and "defend" it in the discussion with other agents (the discussion in which the orchestrators and ultimate evaluators is the LLM). The discussions and the reasoning are logged in an audit trail, which we review on ongoing basis.
- Strategically placed humans in the loop who – depending on the use case – either review AI outputs ex post or act as "human filters" (and accountable individuals from regulatory perspective) who approve the release of AI outputs to the end-users.

Not everything can be disclosed to the end-users for competitive and IP protection reasons, but there needs to be full *internal* transparency of the workings of the system for evaluation reasons.

6. **Make auditability a core feature.**

A core human-centered feature of our systems is end-to-end inspectability: auditors can review the evidence graph and the intermediate execution trace (plans, tool calls, and derived artifacts) rather than only the final narrative. This "glass-box" orientation follows established human–AI interaction guidance that emphasizes visibility into system behaviour and support for effective user control and correction.

The complete execution is logged as an auditable trace: plan steps, tool calls and parameters, generated AEPs, stance edges and their rationales, Planner–Writer exchanges, and the final evidence bundle used for the answer. This directly supports post-hoc audit and targeted failure localization, consistent with recent "evidence pack" agent designs that treat intermediate artifacts and logs as first-class audit objects. It also aligns with record-keeping requirements for high-risk AI systems that emphasize automatic logging to enable traceability across the system lifecycle. Where feasible, the trace supports partial replay (re-executing a subset of steps under the same tool/data/model versions) and structured attestation of actions, consistent with emerging proposals for observability-first agent architectures in high-risk settings.

This auditability-first methodology typically increases latency and compute relative to single-pass generation. However, in accuracy-critical workflows the additional cost is justified when it reduces unsupported claims and enables systematic verification. Iterative "draft–verify–revise" approaches have repeatedly shown measurable hallucination reductions precisely by spending extra steps on verification and revision. In our setting, the same principle is operationalized via explicit evidence acquisition, auditable intermediate artifacts, and human-visible intervention points.

7. **Be fully transparent with the end-users about the constraints and limitations of the system.**

8. **Consider regulatory impacts, which may be very different depending on the exact nature of the use case and data used.**

The key regulation in financial services is MIFID II [11]and its provisions about suitability and client information – and being able to demonstrate that *"natural persons giving investment advice or information about financial instruments, investment services or ancillary services to clients on behalf of the investment firm possess the necessary knowledge and competence to fulfil their obligations"*. This means that any entity providing financial (investment) advice per se without a defensible form of "human in the loop" could fall foul of MIFID II.

Another regulation to consider – although it exists in the UK and a handful of EU jurisdictions, but not on the union level – is SM&CR (senior managers and certification regime) which addresses senior staff's accountability. SM&CR is meant to eradicate "blind spots" and clearly delineate lines of accountability of specific, named senior personnel.

Then there is the EU AI Act: AI systems used in financial services will not fall in the high-risk category (which the AI Act is largely focused on). The EU AI Act also covers what the legislators referred to as "general-purpose AI systems" which LLMs fall under – however, the obligations here are largely on the providers, and also in the contexts that appear to be outside financial services use cases. What is left is the sweeping Article 95, which covers *"drawing up of codes of conduct, including related governance mechanisms, intended to foster the voluntary application to AI systems, other than high-risk AI systems, of some or all of the requirements set out in Chapter III, Section 2*

taking into account the available technical solutions and industry best practices allowing for the application of such requirements".

We interpret Article 95 as a *de facto* requirement to implement most provisions of the act in a flexible, proportionate manner – with national and sectoral regulators deciding whether individual firms have implemented those – technically non-binding – best practices to their satisfaction.

4 Conclusions

Our discussion shows us the financial services industry – using two in-depth use cases – at a crossroad: the promise and the potential of LLMs vs. the huge reputational and regulatory downsides. This, in our view, explains what we consider fairly limited progress in the implementation of AI to date. Also, while financial services industry is heavily regulated, the abovementioned conundrum applies to many other industries; regulated or not.

At the same time, we think it may be somewhat counterproductive to call the current situation an "evaluation crisis" – we think "evaluation challenge" is more suitable. It is also important to acknowledge that barring unforeseen technological revolution that would eliminate LLMs' hallucinations entirely – and some voices in academia say that this is not possible[26, 30]– this is likely a long-term situation.

It is both our firm belief as well as empirical observation that interdisciplinary approach combining expertise in AI, finance, compliance and regulation is the safest and most responsible way to deploy LLMs and other emerging AI technologies[28].

We remain open-minded and enthusiastic about the promises of LLMs, agentic AI and AI in general. We advocate (and practice first-hand in our two AI FinTech startups) a measured, thoughtful, transparent and controlled approach to AI, and particularly to LLMs and other semi-autonomous systems like agentic AI. We hope that our "field tested" observations, reflections and conclusions can be of value to the workshop and its participants.

REFERENCES

[1] Amit Magan, Jean-Manuel Pierron, Jan Goltsche and Benjamin Gauch, 2024. GenAI Lets Telco B2B Sales Teams Get Back to Selling. Boston Consulting Group. https://www.bcg.com/publications/2024/genai-lets-telecom-b2b-sales-teams-get-back-to-selling.

[2] Anne-Laure Mention, 2019. The Future of Fintech. Research-Technology Management, 62(4), 59-63. https://doi.org/10.1080/08956308.2019.1613123.

[3] Bank of England, 2022. Machine Learning in UK Financial Services. https://www.bankofengland.co.uk/Report/2022/machine-learning-in-uk-financial-services. Bank of England.

[4] Bonnie Buchanan 2019. Artificial intelligence in finance.

[5] Bryce Ellencweig, Michael van Oostende and Ricardo Silva, 2024. Gen AI: Opportunities in M&A. McKinsey & Company. https://www.mckinsey.com/capabilities/m-and-a/our-insights/gen-ai-opportunities-in-m-and-a. *Gen AI: Opportunities in M&A*.

[6] Oliver Bussmann. 2017. The Future of Finance: FinTech, Tech Disruption, and Orchestrating Innovation. *Equity Markets in Transition*. R. Francioni and R.A. Schwartz, eds. Springer International Publishing. 473–486.

[7] CFA Institute, 2020. *Artificial Intelligence in Asset Management*.

[8] Claudette Cayrol and Marie-Christine Lagasquie-Schiex, 2005. Graduality in argumentation. Journal of Artificial Intelligence Research, 23, 245-297.

[9] Claudette Cayrol and Marie-Christine Lagasquie-Schiex, 2005. On the acceptability of arguments in bipolar argumentation frameworks. Springer Berlin Heidelberg.

[10] Eduardo Plastino and Mark Purdy, 2018. Game changing value from Artificial Intelligence: eight strategies. Strategy & Leadership, Vol. 46, Iss. 1.

[11] European Parliament, 2014. Directive 2014/65/EU on markets in financial instruments (MiFID II).

[12] Ideagen, 2023. audit fees report. https://staging-corp-waf.ideagen.com/thought-leadership/whitepapers/2023-audit-fees-report.

[13] IOSCO, 2021. The use of artificial intelligence and machine learning by market intermediaries and asset managers. https://www.iosco.org/library/pubdocs/pdf/IOSCOPD658.pdf.

[14] Kate Niederhoffer, Gabriella Rosen Kellerman, Andrew Lee, Anne Liebscher, Kimberly Rapuano and Jeffrey Hancock, 2025. AI-Generated "Workslop" Is Destroying Productivity. Harvard Business Review. https://hbr.org/2025/09/ai-generated-workslop-is-destroying-productivity.

[15] Luisa Kruse, Nico Wunderlich and Roman Beck, 2019. Artificial intelligence for the financial services industry: What challenges organizations to succeed. Proceedings of the 52nd Hawaii International Conference on System Sciences (HICSS).

[16] Michael Chui, Eric Hazan, Roger Roberts, Alex Singla, Kate Smaje, Alexander Sukharevsky, Lareina Yee and Rodney Zemmel, 2023. The economic potential of generative AI: The next productivity frontier. McKinsey & Company. https://www.mckinsey.com/~/media/mckinsey/business%20functions/mckinsey%20digital/our%20insights/the%20economic%20potential%20of%20generative%20ai%20the%20next%20productivity%20frontier/the-economic-potential-of-generative-ai-the-next-productivity-frontier.pdf.

[17] Michelle Lee, Puneet Kher and Gagan Jaggi, 2024. Five priorities for winning with GenAI in wealth and asset management. EY. https://www.ey.com/en_gl/industries/wealth-asset-management/five-priorities-for-winning-with-genai-in-wealth-and-asset-management.

[18] Naren Rama Tadapaneni, 2019. Artificial Intelligence in Finance and Investments. International Journal of Innovative Research in Science, Engineering and Technology, 9(5).

[19] OECD, 2021. Artificial Intelligence, Machine Learning and Big Data in Finance: Opportunities, Challenges, and Implications for Policy Makers.

[20] Office for National Statistics (ONS), 2023. Economic activity and social change in the UK, real-time indicators methodology. https://www.ons.gov.uk/economy/economicoutputandproductivity/output/methodologies/coronavirusandthelatestindicatorsfortheukeconomyandsocietymethodology.

[21] Pankaj Agarwal, Alan Kremer, Ida Kristensen and Audrey Luget, 2024. How generative AI can help banks manage risk and compliance. McKinsey & Company. https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/how-generative-ai-can-help-banks-manage-risk-and-compliance.

[22] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel and Douwe Kiela, 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. arXiv. https://arxiv.org/abs/2005.11401.

[23] Peter Gomber, Robert J. Kauffman, Chris Parker and Bruce W. Weber, 2018. On the Fintech Revolution: Interpreting the Forces of Innovation, Disruption, and Transformation in Financial Services. Journal of Management Information Systems, 35(1), 220-265. https://doi.org/10.1080/07421222.2018.1440766.

[24] Public Company Accounting Oversight Board (PCAOB), 2004. AS 1215: Audit documentation (as amended). https://pcaobus.org/oversight/standards/auditing-standards/details/AS1215.

[25] Puneet Kher and Sameer Gupta, 2025. Unlocking strategic advantage: Generative AI in wealth and asset management. EY. https://www.ey.com/en_us/insights/wealth-asset-management/gen-ai-in-wealth-asset-management-survey.

[26] Soumya Banerjee, Animesh Agarwal and Srijan Singla, 2025. LLMs will always hallucinate, and we need to live with this. Intelligent Systems Conference, Springer Nature Switzerland, 624-648.

[27] Tianyu Gao, Howard Yen, Jiatong Yu and Danqi Chen, 2023. Enabling large language models to generate text with citations. arXiv. https://arxiv.org/abs/2305.14627.

[28] Wojciech Buczynski, Fabio Cuzzolin and Barbara Sahakian, 2021. A review of machine learning experiments in equity investment decision making: why most published research findings do not live up to their promise in real life. International Journal of Data Science and Analytics, vol. 11, 221-242. https://rdcu.be/ch7Xo.

[29] Yu Zhu, Nico Potyka, Daniel Hernández, Yizheng He, Zhanhao Ding, Bo Xiong, Danylo Zhou, Evgeny Kharlamov and Steffen Staab, 2025. ArgRAG: Explainable retrieval augmented generation using quantitative bipolar argumentation. Proceedings of Machine Learning Research, 284, 697–718. https://proceedings.mlr.press/v284/zhu25a.html.

[30] Ziwei Xu, Sanjay Jain and Mohan Kankanhalli, 2024. Hallucination is inevitable: An innate limitation of large language models. arXiv. https://arxiv.org/abs/2401.11817.

[31] *European Parliament and of the Council, 2024. Regulation (EU) 2024/1689 (Artificial Intelligence Act), Article 12 (Record-keeping). Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.*

[32] Hao Chen, Zihan Zhang, Junnan Jiang, Qingjun Su and Qian Xiang, 2026. Holmes: An evidence-grounded LLM agent for auditable DDoS investigation in cloud networks. arXiv. https://doi.org/10.48550/arXiv.2601.14601.

[33] Kevin Wu, Eric Wu, Kerrie Wei, James Zou and Albert Gu, 2025. An automated framework for assessing how well LLMs cite relevant medical references. Nature Communications, 16, 3615. https://doi.org/10.1038/s41467-025-58551-6.

[34] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz and Jason Weston, 2023. Chain-of-Verification reduces hallucination in large language models. arXiv. https://doi.org/10.48550/arXiv.2309.11495.