Financial Conduct Authority

Research Note

30 May 2025

Money Talks: Lessons from Two LLM Pilots on Consumer Guidance

Shuaib Ahmed, Rhosyn Almond, Cameron Belton, Daniel Bogiatzis-Gibbons, Krishane Patel, Manasi Phadnis, Patrick Sholl, Jackie Spang

FCA research notes in financial regulation

The FCA research notes

The FCA is committed to encouraging debate on all aspects of financial regulation and to creating rigorous evidence to support its decision-making. To facilitate this, we publish a series of Research Notes, extending across economics and other disciplines.

The main factor in accepting papers is that they should make substantial contributions to knowledge and understanding of financial regulation. If you want to contribute to this series or comment on these papers, please contact David Stallibrass (david.stallibrass@fca.org.uk).

Disclaimer

Research notes contribute to the work of the FCA by providing rigorous research results and stimulating debate. While they may not necessarily represent the position of the FCA, they are one source of evidence that the FCA may use while discharging its functions and to inform its views. The FCA endeavours to ensure that research outputs are correct, through checks including independent referee reports, but the nature of such research and choice of research methods is a matter for the authors using their expert judgement. To the extent that research notes contain any errors or omissions, they should be attributed to the individual authors, rather than to the FCA.

Authors

The authors were all FCA staff at the time of publication.

Acknowledgements

We would like to thank:

- Cagatay Turkay for some useful early discussions,
- Our other human markers Hammad Bin Hassan, Hassan Tahir, Piotr Piecha, Lesley Dwyer, Taha Mangrio, Glenn Redemann, James Aitken, Richard Boorman, Maria Jomy, Bea Jackson, Hisho Rajanathan, James Macaulay, and Annette Reid for helping us to gather data,
- Grace Elman for supporting the team during the initial setup and prompt engineering phase to ensure the LLM term definition/example generated outputs were collected
- Lawrence Charles, Edmund Towers, Kieran Keohane, and David Stallibrass for managerial support, and
- Internal reviewers for their input and comments.

All our publications are available to download from www.fca.org.uk. If you would like to receive this paper in an alternative format, please call 020 7066 9644 or email publications_graphics@fca.org.uk or write to Editorial and Digital Department, Financial Conduct Authority, 12 Endeavour Square, London E20 1JN.

Contents

Introduction to the Two Pilots

Artificial Intelligence (AI) has the potential to drive inclusive innovation and growth in financial services through improved consumer support. However, to meet the full potential of AI in retail financial services requires adequate testing to ensure sufficient mitigations are included. This paper is part of a set of Research Notes which set out to inform public discussion and provide rigorous research on topics that relate to AI.

This note explores the potential usefulness and limitations of Large Language Models (LLMs) (the generative AI such as the GPT series of OpenAI) in consumer-facing contexts in financial services. This is not yet well established – though there are some studies in other contexts, such as increasing productivity by consultants (Dell'Acqua et al. 2023) but decreasing the effectiveness of searching for information in a public sector context (Behavioural Insights Team 2023). We decided to explore how to test generative AI and its applications to consumer financial services in two pilot projects:

- In the first project, we examined simplifying complex financial concepts through asking GPT-3.5 and GPT-4 to provide lower reading age definitions with appropriate examples.
- In the second project, we examined providing guidance using LLM-generated content in a chatbot with predetermined questions and answers in a cash savings context, as compared to using a website-based Q&A page.

We undertook this research both to examine the potential methodologies that can be used by firms to research consumer outcomes from the use of LLMs, and to add to our understanding of this evolving technology.

The FCA's approach to the use of AI by the firms we regulate is outcomes-focused, and therefore, testing – utilising a variety of methods – can be an important way for firms to understand the outcomes their consumers are receiving. This can help them to identify potential issues and provide an inclusive approach, which can support consumer trust. Trust contributes to positive interactions with consumers, and can lead to better consumer outcomes where it promotes, for example, saving.

We have recently published an engagement paper on a proposal for pilots for testing live AI models (FCA, 2025), which aim to support the safe and responsible deployment of AI by firms and achieve positive outcomes for UK consumers and markets.

Given the focus and aims of the work presented readers should be aware that this note:

- constitutes research to spark debate and contribute to academic discussion, rather than any form of guidance or direction about what firms or practitioners should do;
- does not set out any expectations for how firms should approach managing AI risks (in all cases, firms will need to consider the risks relating to AI adoption in the context of their specific use cases and considering applicable requirements).

Why test LLMs?

Financial decision-making involves finding and processing lots of information

Retail financial services can be complex to navigate for consumers and making sensible product choices in an informed way is a critical component of that journey. Behavioural science tells us that humans have limited cognitive capacity, which restricts our ability to always be fully rational or perform complex computations for decision-making. With a vast number of financial decisions to make and a wide variety of financial products to choose between, finding the right information can be difficult. When it takes time to find information, this can put people off from searching for that information, even if this causes negative economic consequences (Hsiao et. al. 2021).

Behavioural science also tells us that consumers are subject to several common behavioural biases, particularly in cases where there is lots of information and complex decision-making. Some examples of these include:

- **Status-quo bias**, where consumers often remain with their current product, even if there are better products available to them (see e.g. Eidelman and Crandall 2012).
- **Present bias**, where consumers overweight the immediate costs or benefits of a product compared to their long-term value (see e.g. O'Donoghue & Rabin, (1999)).
- **Choice overload**, where consumers may find having too many options and variation of products difficult and taxing to understand (see e.g. Chernev, Böckenholt, and Goodman 2015).

Enter the LLMs

With these behavioural challenges in mind, there could be a role for LLMs to help humans find and process these vast amounts of information to help make more informed financial decisions. For example, LLMs could quickly provide definitions of any confusing terms in a timely manner, as consumers search for information. This could improve information "retrieval" – while LLMs do not directly search for documents like search engines, LLMs can draw upon the documents they are trained on to "find" facts and related concepts with which to explain terms. This can be quite sensitive to the phrasing of the question asked as a result. Moreover, LLMs can deploy helpful language like metaphors and similes while answering human queries (Ichien, Stamenković, and Holyoak 2024). These could support consumers in thinking about their finances in ways that have not been fully explored before.

Some initial evidence shows that LLMs can be used to generate suitable financial recommendations for investing, and that historical performance of the recommended portfolios meets benchmarks for professionally-managed portfolios (Fieberg et al., 2024). LLMs have also been shown to enhance productivity among management consultants (Dell'Acqua et al. 2023) and to create timely feedback to enhance learning (Meyer et al. 2024).

However, LLMs can also make mistakes in explaining concepts because they may struggle to reason logically (Arkoudas 2023), lack planning in incremental tasks (Bubeck et. al. 2023) and can produce societally harmful content (OpenAI et. al. 2023; for a taxonomy of risk see Weidinger et. al. 2022). Further, LLMs may exhibit some of the same biases as humans, as they ultimately do learn from human text (Ross 2024, Zhou et al. 2024, Dwyer et al., 2025).

One concern, which we do not explore here, is that LLMs could cause consumers to ask for financial advice less, even on consequential matters, or to seek out other informative guidance less.

Our focus

In this paper, we concentrate on the tasks of producing simple definitions of complex financial terms (such as for compound interest) and guidance (understanding features of one type of cash savings product). We recognise that these are only two of many use cases. However, simplification was chosen as a focus given the low level of financial literacy and numeracy among the general UK population (MaPs 2021).

Previously, work at the FCA by Chak et al. (2022) showed that a simple automated 'robo-advice' tool significantly improved borrower repayment decisions in a randomised controlled trial.

While LLMs could be part of the picture in improving customer journeys in financial services, this requires careful technical and consumer testing in contexts where they may be used. We summarise two pilot projects here and set out the detailed methodology and results in the attached Annexes for both projects. The three main lessons we learned from running two pilots were:

1. LLMs do seem to be able to accurately simplify concepts and make text easier to read, but comprehensively testing their outputs requires both human and automated evaluation.
2. Specific content and information presentation - including how and where an LLM is used and incorporated within a customer journey - are likely to affect outcomes like consumer comprehension and engagement with guidance.
3. Many consumers seem keen for some type of automated support.

Lesson 1: LLMs do seem to accurately simplify concepts and reduce the reading level, but comprehensively testing their outputs requires both expert human and automated evaluation.

One pilot involved simplifying financial concepts, by giving the task of producing appropriate definitions and examples of common financial terms (like interest rates, arbitrage, and annuities) to two versions of OpenAI's GPT family of LLMs. We sourced the terms from the Plain English Campaign (2024) for personal finance concepts, and the Federal Reserve Bank of St Louis focussing on economics-relevant financial concepts. We ensured that any American-specific terms were removed from consideration.

LLMs take prompts, which can be divided into overall or "system" prompts and more specific task-based prompts. In the "system" prompt, we asked the LLM to adopt the persona of a financial literacy teacher and expert in finance and economics, not provide financial advice, to write as if to an 8-year-old living in the UK and to give simple but accurate definitions with an appropriate example. One problem this created was the use of childlike imagery, which remained even when asking the LLM to write at an 8-year-old reading level or just a simple reading level. This could be important to avoid as it may reduce trust and increase confusion.

We then varied the degree of instruction to the LLM in the task-based prompts, either giving it a minimal instruction (naïve prompting), more specific instructions (zero-shot prompting) or a structure to respond with (few-shot prompting). More details on the prompts used are in the Annexes to this note. That said, perhaps because of the relative simplicity of the task, we saw little evidence of differences in the human or automated evaluation between either GPT3.5 or GPT4, and between increasingly sophisticated prompting. That could be because producing simplified financial concepts is a reasonably simple task for an LLM.

We double-marked the responses (after providing initial guidance on the meaning of the scales and the task for consistency) using expert human evaluators (FCA staff) who rated the responses on a 1 to 5 scale for each of:

- accuracy (5 completely accurate to 1 entirely inaccurate),
- absence of misinformation (5 completely absent to 1 very serious misinformation),
- absence of jargon (to proxy for readability, 5 entirely jargon-free to 1 overwhelming use of jargon),
- and appropriateness of example (5 entirely appropriate example to 1 no or inappropriate example).

We also used automated methods, based on simple algorithms or formulae, to evaluate the readability and semantic similarity (i.e. having a similar meaning) to the original definitions to proxy accuracy.

We found that overall, the responses seem high-quality: over 90% of responses on both models obtained a score of 4 or 5 out of 5 when rated by FCA staff on all the criteria. We saw a marked reduction in the reading age required. The human-written definitions were at a 9th/10th grade (14–16-year-old) reading level but the best LLM responses were at a 4th-6th grade (9–10-year-old) reading level. This is not quite like writing for an 8-year-old reading level, but these are substantial improvements. Note that American year groupings are used as this is standard in reading age measurement.

However, the human and automated methods had mixed agreement on whether the definitions of specific terms were accurate and are complements to each other rather than substitutes. For example, expert human review can help best identify the technical components of whether definitions are correct, whereas automated methods might be better placed to answer questions around reading levels. Why are they complements rather than substitutes?

First, there is a difference in what expert humans and automated methods can measure. Expert human markers can examine the quality of the definition more directly than automated methods. In particular, if the LLM improves on a human-written definition, an expert human marker would give it a high score but our automated similarity metric would mark it down for being dissimilar. Our automated methods look at word and sentence complexity to measure readability, and they can be used to compare different LLM models and prompting methods in terms of how well they replicate the content of the original definitions. This is likely to be a better proxy than the judgment of expert human markers on how jargon-free text is because expert human markers are (typically) more educated and technically knowledgeable than the general reading public. That said, a true measure of how readable text is would require direct testing with consumers.

Secondly, automated, and human marking would be substitutes only if similar measures were highly correlated with each other. For example, we might have expected to see that highly readable responses would score well overall, or that highly rated responses (especially for accuracy) would have high semantic similarity (which measures if the LLM definitions have a similar meaning to the reference definitions). However, we saw very little correlation between our human metrics and our automated ones. We did see some association between accuracy and semantic similarity – but this may be because we also gave the expert human markers the reference definitions as an indication of what good might look like.

Third, we studied cases where the human and automated methods disagreed to try to understand why. From that, we make some tentative hypotheses:

- LLMs might have produced good new definitions that were phrased in quite different ways to the original definitions, meaning they had high human ratings but low semantic similarity.
- However, expert human evaluators may still rate more complex answers highly, given that precise accuracy might be more important to them than the average consumer.
- LLMs did sometimes generate childish examples that are inappropriate. This may reduce its human scoring in other areas, despite potentially maintaining high semantic similarity.
- There were some concerns with readable text generated by LLMs being inaccurate or almost misinformation, though there were relatively few examples of this.

It is worth noting to conclude, that neither expert human marking nor automated methods are a perfect substitute for testing with consumers, however, they are likely to be cheaper and more practical in certain low-risk instances and are a helpful first stage in testing.

Lesson 2: Testing a limited version of LLM-generated guidance with consumers shows that *how* LLMs are rolled out matters

In our second pilot, we created a customer journey simulation of a savings account selection so that we could look at behaviour in context. This allowed us to observe how consumers might interact with LLM generated content, presented through a limited chatbot interface, where they could only select from pre-defined questions to ask the chatbot.

Our study varied the context in which financial guidance was presented: Q&A accordion style (a format used to present questions and answers in a way that allows users to click on a question to expand and reveal the answer beneath it) and chatbot interface, and the content of the guidance itself – human written vs LLM-generated.

We tested across three treatment/control groups: Q&A only; Q&A and chatbot; and chatbot only.

We found that:

1. Adding a chatbot onto the Q&A guidance did not help participants choose the right account. Those in the chatbot only group were worse at choosing the right savings account compared to the Q&A only and Q&A and chatbot groups. This indicates that even though LLM content has the potential to be useful, how the content is presented to consumers might affect how useful it actually will be.
2. Engagement with the chatbot was low, so people saw less information in the chatbot group. Even if they had been useful, consumers may not use chatbots

that are made available to them. The limitations of our chatbot – participants could not type their own question but had to select them from a predetermined set - may have driven low engagement.

3. However, even highly engaged participants who launched and interacted with the chatbot were less likely to select the right account compared to the Q&A groups. This implies that this result was driven at least in part by differences in how information was presented. This could have affected how participants interacted with and understood information provided by the chatbot versus the Q&A.
4. By contrast, financial comprehension was not strongly affected by seeing either the Q&A page or the chatbot.

Finally, those in the groups with a chatbot were more likely to report that they would use one for financial decision-making in future, despite these groups having lower average performance in the savings account selection task. This demonstrates that relying on attitudinal survey data may not provide the full picture of consumer-AI interaction, and it is important to measure behaviour directly.

This is similar to the picture in Belton et al. (2025), where we found that consumers had increased confidence in their ability to challenge an incorrect AI credit decision when given more complex explanations, but their actual performance in the task was sometimes worse. Positive sentiment alone does not necessarily correlate with improved behavioural outcomes.

It is important to highlight that our results are highly context-specific and exploratory; we cannot presume what has occurred in one context will work in another. The chatbot interface had significant limitations, and a more dynamic version could have led to different results. Our work highlights how seemingly minor changes can influence customer understanding and product selection, highlighting the need for careful testing. The more interactive elements of a LLM interface like ChatGPT might be more effective in certain respects than static chatbots, though they are more difficult to control the outputs of, so come with their own risks.

Further work could look to dig deeper and look to isolate the effect of just the language, or presentation, or the context - using our results to iteratively test and learn to optimise outcomes.

Lesson 3: Many consumers seem keen for some type of automated support

We asked consumers two questions at the end of the survey relating to the second pilot:

1. “How might you use a chatbot to help with making decisions about your finances or to gather information about financial products?”
2. “What would be the first question you ask a chatbot, if you were to use it to help you make decisions about your finances or to gather information about financial products?”

Overall, we found that:

a. Across the different treatment groups, respondents were interested in comparing products (focusing on strategies to grow wealth and maximise returns), understanding financial jargon, and seeking personalised advice on savings and investments (noting that this was likely influenced by the fact that they had just seen a problem related to savings).

b. Shared questions included finding the best interest rates for savings accounts and ISAs, as well as comparing products for specific needs. All groups were interested in selecting the best savings accounts based on personal needs, with a focus on both short and long-term saving strategies.

c. Which treatment the participants saw affected their responses. For example, participants exposed to the chatbot tended to prioritise immediate decision-making and resource gathering.

In summary, these free-text responses did seem to indicate a willingness to use chatbots in helping make decisions or gather information, including around complex questions on savings strategies.

In our pilot, we found that exposing people to a LLM chatbot meant they were more likely to want to use AI to help them make financial decisions in the future. These results align with our past research on robo-advice, which showed that the average willingness to pay for robo-advice exceeded its monetary benefits (Chak et al., 2022). This research suggests that consumers might value non-monetary benefits of automated support, such as avoiding the cognitive and psychological costs of making choices that are consequential to their financial lives.

What's next? Some suggestions for future research

Our lessons show that there are important opportunities to provide effective and tailored support with AI. For example, this could include providing financial guidance, tailored explanations of financial concepts, or agent-based chatbots for resolving complex consumer queries. However, they also show that careful testing is important and that we still have lots to learn about how to use AI with consumers.

The FCA will be conducting pilots with firms of AI models and remains engaged with the idea of testing. Further, we would welcome potential pilots on questions similar to those engaged with in this note.

In terms of future research, one clear avenue to pursue is further testing to better understand how LLMs can be designed and integrated into consumer journeys such that the potential benefits, such as timely information simplification, can be fully realised. Another avenue is to explore more complex use cases and for academics to continue to test state-of-the-art models in specific industry contexts.

Further research could explore newer evaluation methods such as LLM-as-a-judge (see e.g. Gu et al. 2024), which involves testing LLM outputs using a second LLM to evaluate performance or other indicators of quality. It could be interesting to study if our finding of low association between human and automated evaluation methods persists for LLM-as-a-judge. That said, LLM-as-a-judge itself would need to be validated in this context and is unlikely to be a replacement for at least some level of manual review of outputs.

Finally, in terms of future experimental testing with consumers, one could:

- Examine the use of LLMs in a more interactive setting or live interaction with LLMs, rather than a fixed pre-determined chatbot.

- Analyse the effect of personalised personas, which might both make information seem more appealing but create higher risks of being treated like an actual financial advisor.
- Expand the scope of the industries studied beyond savings products, or to specific customer bases which might have different access or vulnerability needs.

12 Endeavour Square, London E20 1JN
Telephone: +44 (0)20 7066 1000 Website: www.fca.org.uk