Financial Conduct Authority

# Annexes

30 May 2025

## Money talks: Lessons from Two Pilots into LLMs for financial guidance

**Shuaib Ahmed, Rhosyn Almond, Cameron Belton, Dan Gibbons, Krishane Patel, Manasi Phadnis, Patrick Sholl, Jackie Spang**

# Contents

# Annex 1: Simplification Pilot

## Prompts for Simple Definitions Task

We set out to research how automated and human methods might be used to evaluate new use cases for LLMs, settling on a simplifying definitions task as previously explained. For this task, we used 241 terms, relating to basic finance and economics concepts, sourced including their reference definitions from the Plain English Campaign (2024) and the Federal Reserve Bank Of St. Louis (2024).

As with many other implementations of LLMs, GPT-4 (we use gpt-4-1106-preview, which was last updated November 2023 and includes knowledge up to April 2023) and GPT-3.5 (we use gpt-3.5-turbo-1106, with a knowledge cut-off around September 2021), contain numerous defences or "guardrails" against harmful content, including self-harm, hate, pornographic, harassing, demeaning, violent, or illegal content (see e.g., OpenAI 2023).

However, through careful prompting, LLMs can also learn other rules that developers might want to enforce, even up to and including more qualitative norms like replying with a particular complexity or forbidding it giving out certain policy information or certain kinds of legal guarantees. We therefore tested these two models under naïve (where the LLM saw a very basic prompt without much context), zero-shot (where the LLM saw a more context-specific prompt), and few-shot (where in addition to the zero-shot prompt, the LLM also saw 20 of the reference examples).

We aimed to choose prompts that asked the GPT models to produce simple definitions of complex financial concepts, and to see how these are evaluated by both automated methods and human reviewers. Ideally, automated and human reviews would be strongly positively associated with each other. Then automated reviews might be able to be used in a higher volume of testing because they are cost-effective and scale more easily as humans can only read so much text in a short space of time. We find that…

Our prompts contained the following elements suggested by White et. al. (2023):

- Asking the model to adopt a persona, which has been suggested to be useful when seeking to have an LLM "retrieve" or more accurately imitate elements of specialised knowledge and official tone or style.
- Managing the context of the response, such as asking the LLM not to contravene certain regulations, or to include particular times, places, or kinds of information in its response.
- Asking the LLM to ensure its own responses are accurate by prompting the LLM to check the assumptions behind its answers and identify any errors.
- Asking the LLM to provide its answers in a particular structure and giving a template for that structure. This ensures that information is logically presented and means that less post-processing is required to have readable answers.

We opted for responses with a low degree of random word generation, which might be less appropriate for prompts to ask for brainstorming creative ideas but seemed sensible given the formal and logical nature of the responses we are looking for. To do that, we always set GPT's temperature to zero which produces more conservative and predictable responses, and its "Top P" to 0.01 which means it uses a less wide-ranging set of words. In exploring responses where we alter these parameters, they seem to be on the face of it strictly worse, especially for high temperatures producing entirely nonsense responses.

Finally, given low general functional literacy and numeracy levels, as well as a need to ensure it gives answers in a UK context, we added that the LLM should give answers "as if you were talking to an 8-year-old who lives in the United Kingdom". One side effect of this is that apart from just simplifying the language or concepts it uses, the LLM began to sometimes deploy childlike imagery in its definitions. For instance, it explained compound interest through the idea of multiplying money fruit. Other examples of childlike imagery include references to money trees, piggy banks, pocket money, toys, parents, and lemonade stands. This occurs for both definitions and examples. This tended to remain even if the prompt was refined to say that the LLM should write at an 8-year-old's reading level, or just a simple reading level – the ask for simple language seems to prompt childlike imagery or language regardless.

Sometimes, that had the ironic effect of making the definition more complicated to understand, by defining an annuity as 'like' a birthday cake distributed over time, someone needs to both understand the analogy and translate it back into money terms. The use of childlike imagery or language can make the definition incorrect, for example by implying money from a debit card is 'magic', when in fact it must come from wages or other sources of earnings. Similarly, with the 'money fruit' analogy, compound interest does lead to the accumulation of money over time, but it does not do so at a rate or through a process comparable to replanting crops.

The exact prompts used were:

- System/model prompt – a prompt that sets out the intended guardrails/constraints of the LLM to limit the chance of unintended responses/outputs. System prompts are present by default beforehand when attempting the text input prompts as seen below

*You are an experienced financial literacy teacher and an expert in finance and economics. Provide a simple, but accurate definition of financial technical jargon as if you were talking to an 8 year-old who lives in the United Kingdom. You are not a financial advisor so you must not provide financial advice under any circumstances, complying with Financial Conduct Authority regulations. Prioritise the following when formulating your response: 1) do not give financial advice, 2) ensure the information provided is accurate and relevant to a financial context, and 3) maintain simplicity. For each financial term, provide a financial definition and example using the following format:*

*Term: …*
*Definition: …*
*Example: …*

Text input prompts (prompt engineering) - when considering the quality of output/responses, there are known techniques to get a specific response(s). The **three** prompts we use are non-exhaustive but are core to various LLM applications in terms of finding the optimal combination given the use case or task to be solved. To reduce manual effort, we increased the sample size within the LLM input dialogue to circa 20 terms per prompt generation.

- Naïve prompt - passing the term to the user prompt with minimal instruction, which can be seen as a baseline.

*Explain the following term(s):*

*"""*
*{Terms are passed within the delimiters or this can be collapsed into a singleton}*
*"""*

Zero-shot learning prompting - When the term is passed as input to the prompt, it also follows with some basic direction to guide its response

*Simplify, summarise and provide a definition for each of the following financial term(s):*

*"""*
*{Terms are passed within the delimiters or this can be collapsed into a singleton}*
*"""*

Few-shot learning prompting - When the term is passed as input to the prompt, it follows some direction and instances of desired terms along with their corresponding definitions and examples, exploiting the technique of 'in-context learning' to guide its response.

*Simplify, summarise and provide a definition for each of the following financial terms:*

*"""*
*Term: Allocation rate*
*Definition: This tells you how much of the money you put into something like a savings plan actually goes into the investment after fees are taken out.*
*Example: If you put £100 into a savings plan and the allocation rate is 90%, then £90 is actually added to your plan and the rest might be used to pay for the plan's costs.*

*Term: APR*
*Definition: APR stands for Annual Percentage Rate, and it tells you how much it costs to borrow money for a year, including fees and interest.*
*Example: If you borrow £100 with an APR of 10%, you'll have to pay back £110 after one year.*

*Term: Bank statement*

*Definition: This is a note from the bank that tells you all the money that went in and out of your account over a certain time.*
*Example: It's like getting a monthly report of what you've saved and spent from your piggy bank.*

*Term: Bond*
*Definition: A bond is like a loan you give to a company or the government, and they promise to pay you back with a little extra after some time.*
*Example: If you buy a bond from the government, they might use your money to build schools, and later they'll give you your money back plus some more as a thank you.*

*Term: Borrowing*
*Definition: Borrowing is when you get something, like money, from someone else with the promise to give it back later, usually with a little extra as a thank you for letting you use it.*
*Example: If you borrow £10 from your friend to buy a toy, you might give them £11 back next week as a thank you.*

*Term: Capital adequacy requirement*
*Definition: This is a rule that says banks must keep a certain amount of money safe in case something goes wrong and they need it.*
*Example: It's like having to save some of your pocket money instead of spending it all, so you have some left over for emergencies.*

*Term: Capital Investment*
*Definition: Capital investment is when you spend money on things that can help make more money in the future, like machines for a factory or a new shop.*
*Example: Buying a new ice cream maker for your ice cream stand so you can sell more ice cream is a capital investment.*

*Term: Credit agreement*
*Definition: This is a formal agreement where someone agrees to lend you money and you agree to pay it back, usually with extra (interest).*
*Example: If you get a loan from the bank to buy a computer, you have a credit agreement with the bank.*

*Term: Diversification*
*Definition: Diversification is when you spread your money across different things instead of just one, so if one doesn't do well, you don't lose all your money.*
*Example: Instead of buying just one kind of toy, you buy different kinds, so if one breaks, you still have others to play with.*

*Term: Equity*
*Definition: Equity is the part of something you own that is really yours. If you buy something like a house with help from a bank loan, equity is the part of the house's value that you've paid for yourself.*

*Example: If you buy a toy for £10 and you paid £6 of your own money and borrowed £4, your equity in the toy is £6.*

*Term: Facility*
*Definition: In finance, a facility is a special kind of agreement, like a promise from a bank that they will lend you money when you need it.*
*Example: It's like having a special deal with your school canteen that you can get a snack even when you don't have money with you, and you pay them back later.*

*Term: Securities*
*Definition: Things like stocks or bonds that people can buy and sell as a way to invest money.*
*Example: It's like buying a tiny piece of a company or lending money to someone, and later they give you back more.*

*Term: Spot rate*
*Definition: The price of something like money from another country if you want to buy it right now.*
*Example: If you want to buy American dollars with your British pounds today, the spot rate tells you how much it costs.*

*Term: Standing order*
*Definition: Telling your bank to pay the same amount of money to someone else regularly without you doing anything.*
*Example: Imagine if your piggy bank automatically gave money to your friend every week for their comic book.*

*Term: Stockbroker*
*Definition: A stockbroker is a person whose job is to buy and sell stocks for other people.*
*Example: If you wanted to sell some of your toys, a stockbroker would be like a friend who helps you find someone to buy them.*

*Term: Surety*
*Definition: Surety is when someone promises to pay money if another person can't pay back what they owe.*
*Example: If you borrow a toy from a friend and your big brother says he'll give one of his toys if you don't give the toy back, your brother is being a surety.*

*Term: Tax allowance*
*Definition: A tax allowance is an amount of money you can earn without having to pay tax on it.*
*Example: It's like if you're allowed to earn £10 from doing chores before your parents ask you to contribute to the family piggy bank.*

*Term: Term loan*
*Definition: A term loan is when you borrow money from a bank and agree to pay it back with a little bit extra (interest) over a set period of time.*

*Example: Your mum borrows £1,000 from the bank to buy a new oven and agrees to pay back £1,100 over 2 years.*

*Term: Transferable securities*
*Definition: Transferable securities are like special tickets (stocks, bonds) that you can buy, sell, or give to someone else, and they can be worth more or less money over time.*
*Example: If you have a Pokémon card that you can trade with others and its value can go up or down, it's like a transferable security.*

*Term: Warrant*
*Definition: A warrant is like a special ticket that can let you buy something in the future, like a company's stock, at a set price, even if the price goes up or down.*
*Example: If you have a coupon that lets you buy an ice cream for £1 anytime in the next month, even if the price goes up, it's like a warrant.*
"""

*Please simplify, summarise and provide a definition for the following financial terms:*

"""
*{Terms are passed within the delimiters or this can be collapsed into a singleton}*
"""

## Automated Evaluation: Design and Results

Automated Evaluation was conducted to see both if the generated definitions were simpler (as measured by improvements in their readability) while maintaining their meaning (as measured by their similarity to given reference definitions). The measures we use here are, of course, much simpler to generate than human evaluations but equally only proxy for these concepts because of the nuances of language use.

To measure readability, we chose the Dale-Chall Readability Index (1948), which has lower scores for simpler or more readable texts, when sentences are shorter and composed of fewer difficult words (those not in a given list of simple English words, or their plurals or derived verb forms). An index of 4.9 or lower is considered easily understood by the average 4$^{th}$ grade student in a US context, 5.0-5.9 (5$^{th}$/6$^{th}$ grade), 6.0-6.9 (7$^{th}$/8$^{th}$ grade), 7.0-7.9 (9$^{th}$/10$^{th}$ grade), 8.0-8.9 (11$^{th}$/12$^{th}$ grade), 9.0-9.9 (university student).

To measure the semantic similarity between generated and reference definitions, one needs to overcome the fact that very different word choices can express the same idea. Therefore, we used two measures based on pre-trained *embeddings*, which look at words which occur together in the same texts, and then attempt to quantify how related they are, for instance the word "cat" and "dog" might be quite semantically similar as they are both nouns which refer to pets. As they are based on sets of documents which have biases on the basis of protected characteristics like gender and race, the use of embeddings is not without its problems, as we have previously documented in Dwyer, Francis, and Tyagi (2024). However, in this circumstance, we consider the risk of bias to be low for two primary reasons: the definitions are primarily measuring abstract concepts that do not

refer to the characteristics of people and the results are being used to understand the performance of LLMs not make decisions about any real persons.

We then used the SentenceTransformers embeddings (Reimers and Gurevych, 2019; Reimers and Gurevych, 2020), which is based on a variety of sources crawled from the Internet. In particular, we used their implementations on Hugging Face, available at https://huggingface.co/sentence-transformers/all-mpnet-base-v2 respectively.

In both cases, we compute the cosine similarity between the generated and reference definitions, which ranges from 0 representing entirely unrelated meanings to 1 meaning perfectly similar meanings.

Average results for semantic similarity and readability are in Table 1, where we find that:

- Semantic similarity is similar across different models and prompting methods.
- Semantic similarity measures are more appropriate for ranked comparisons and do not allow for easy conclusions around whether the definitions are appropriately similar to the reference definitions.
- There are substantial improvements in readability between GPT-3.5 and GPT-4, though both do improve readability.
- The reference definitions have a reading age of a Grade 9 or 10 student (14-16 years old) whereas the best definitions from GPT-4 (zero-shot and few-shot) have a reading age of a Grade 4 student (9-10 years old).

In Table 2, we analyse the effects of prompting and of using either GPT-3.5 and GPT-4 on both semantic similarity measures and Dale-Chall readability using linear regressions with random effects for the words. We find GPT-4, surprisingly, does worse across the two measures, though these effects are small, even if statistically significant. We see that zero-shot does better than naïve prompting in terms of our mpnet-base-v2 measure of semantic similarity, surprisingly showing a larger effect than few-shot, even if the effect again is small.

### Table 1: Average semantic similarity and readability by model/prompting combination

| Data unit | SentenceTransformers Semantic Similarity (mpnet-base-v2) | SentenceTransformers Semantic Similarity (MiniLM-L6-v2) | Dale-Chall Readability (reading complexity) | Reading Complexity Difference |
|---|---|---|---|---|
| **Naïve GPT-3.5** | 0.66 | 0.62 | 7.00 ($9^{th}$-$10^{th}$ grade) | -0.52 |
| **Zero-shot GPT-3.5** | 0.65 | 0.62 | 6.84 ($6^{th}$-$9^{th}$ grade) | -0.68 |
| **Few-shot GPT-3.5** | 0.67 | 0.63 | 6.58 (6th-9th grade) | -0.94 |
| **Naïve GPT-4** | 0.65 | 0.61 | 5.20 ($4^{th}$-$6^{th}$ grade) | -2.33 |
| **Zero-shot GPT-4** | 0.64 | 0.6 | 4.97 ($4^{th}$-$6^{th}$ grade) | -2.55 |

| | | | | |
|---|---|---|---|---|
| **Few-shot GPT-4** | 0.66 | 0.62 | 4.99 ($4^{th}$-$6^{th}$ grade) | -2.53 |

**The first three columns give semantic similarity in terms of average cosine distance for the two sets of embeddings (scale 0 to 1) and the second last column gives the average readability score (the reference definitions have an average Dale-Chall readability of 7.52 corresponding to a reading age of 14-16 years old ($9^{th}$-$10^{th}$ grade) and its grade equivalent in US schooling terms in brackets. The last column shows the (average) reading complexity difference, remember negative is an improvement, based on average readability for reference definitions.**

### Table 2: Summary of regression results for semantic similarity and readability

| Data unit | SentenceTransformers Semantic Similarity (mpnet-base-v2) | Dale-Chall Readability Score Change Reference over Definition |
|---|---|---|
| **GPT-4** | -0.014***<br>(0.003) | -1.751***<br>(0.105) |
| **Zero-shot** | 0.008*<br>(0.004) | -0.190<br>(0.128) |
| **Few-shot** | 0.003<br>(0.004) | -0.04019 |

**The first three rows give the regression coefficients across the four models, where GPT-3.5 and naïve are the omitted dummy variables. Semantic similarity is bounded by 0 and 1, and in the final column they are in terms of changes in average Dale-Chall Readability Index.**

## Human Evaluation: Design and Results

Human evaluation was conducted primarily to ask whether the LLM outputs were accurate, as automated methods might fail to notice nuances of meaning that make a definition mistaken. Raters marked each definition/example pair on the following five-point Likert scales:

a. Accuracy of definition (5 most accurate).
b. Absence of misinformation (5 no misinformation, 1 contains significant misrepresentations). We included this outcome separately because we wanted to differentiate between cases where a definition is incorrect because it excludes key information and when it includes incorrect information.
c. Jargon (5 most jargon-free). We included this because the task is around simplification – so if the definitions were accurate but contained jargon, they could not be considered simplified. This should complement automated readability metrics.
d. Appropriateness of example did the explain adequately explain the concept (5 explains concept comprehensively).

Each model/shot combination (e.g., GPT-3.5 zero-shot) had two human reviewers, who were all FCA staff members from relevant functions (Supervision including Asset Management, Mortgages, and Consumer Investment, Authorisations, Economics, and Advanced Analytics – our data science function). One advantage of using FCA staff is that they have knowledge of financial markets needed to make these assessments, but it may mean their judgments were less representative of the population – particularly in what

they view as jargonistic or helpful in an example. Even with this relatively simple task, it is worth noting that several evaluators noted that it was difficult to provide consistent, reasoned scores, often requiring reflection to tweak the scores, and it was quite subjective.

In general, the evaluators found the performance of the LLMs at providing accurate simplified definitions and examples to be high, as seen in Table 3.

### Table 3: Percentage of 4 and 5 ratings by outcome and model

| Data unit | Accuracy | Absence of misinformation | Absence of Jargon | Appropriateness of Example |
|---|---|---|---|---|
| **Naïve GPT-3.5** | 93.5 | 96.9 | 72.9 | 93.9 |
| **Zero-shot GPT-3.5** | 96.3 | 93.8 | 92.1 | 98.3 |
| **Few-shot GPT-3.5** | 92.9 | 96.5 | 83.6 | 95.0 |
| **Naïve GPT-4** | 94.4 | 100 | 94.4 | 98.3 |
| **Zero-shot GPT-4** | 82.2 | 95.8 | 86.5 | 91.9 |
| **Few-shot GPT-4** | 92.5 | 96.3 | 88.6 | 88.8 |

**Proportions include, in the vast majority of cases, scores from two reviewers for each term. There were a small number of terms that were not answered by one reviewer.**

We then analyse whether there are differences between GPT-3.5 and GPT-4 (using GPT-3.5 as the base case), as well as whether there are differences between naïve, zero-shot, and few-shot prompting (use naïve as the base case).

We use logistic regressions (of which the results are in Table 4) on a binary outcome of a 4 or 5 rating being counted as a 1 ('success', as we want the models to have a high degree of accuracy and appropriateness) and a rating of 1, 2, or 3 being a 0 ('failure') with random intercepts for the raters (to control for differences between raters) and the words. We note that this analytical strategy was not pre-specified, and these results should be taken as exploratory.

We present average marginal effects in Table 4, which can be interpreted as the percentage point change in probability of success relative to the base case. For example, a marginal effect of 0.100 would indicate a 10-percentage point increase in the probability of success compared to the base case. The binary outcome corresponds to the high quality we would ultimately desire in language model outputs for consumer applications.

### Table 4: Summary of regression results for manual marking

| Data unit | Accuracy | Absence of misinformation | Absence of Jargon | Appropriateness of Example |
|---|---|---|---|---|
| **GPT-4** | -0.038<br>(0.029) | 0.026<br>(0.021) | 0.041<br>(0.044) | -0.003<br>(0.037) |
| **Zero-shot** | -0.031<br>(0.035) | -0.047<br>(0.024) | 0.045<br>(0.054) | 0.014<br>(0.033) |
| **Few-shot** | -0.007<br>(0.033) | -0.037<br>(0.026) | 0.015<br>(0.062) | -0.047<br>(0.056) |

**The first three rows give the regression coefficients across the four models with average marginal effects provided, with standard errors in parentheses (* p < 0.05, ** p < 0.01, *** p < 0.001).**

Where the models did make mistakes, for definitions one of the following were usually true:

- It missed out parts of the definition, for example, in economics or finance a good is not just a product but something that satisfies a particular want, such as food satisfying hunger.
- It missed the financial context e.g., 'contracting in' is a term specifically concerning pensions, but the given definition for 3.5-zero-shot did not mention that.
- It picked overly narrow contexts e.g., the definition for direct deposit for 3.5-zero-shot makes it sound like direct deposits are only used for wage payments, but of course that is only one use for them.
- The definition requires knowledge of other financial terms (e.g., requiring an understanding of balance sheets, without also defining them, in order to define a maturity).

By contrast, where generated examples were not appropriate, this often occurred because they falsely implied circumstances that occasionally occur always happen. For instance, a car loan going into default requiring the seizure of the car (which may or may not occur). Examples sometimes also failed to explain why something happens e.g., the example given for depreciation on the 3.5-zero-shot model just says that driving a car out of the lot depreciates its value. This is both a somewhat odd example (a more typical would be that cars depreciate over time because of wear-and-tear) and doesn't explain why it happens (customers want new not even slightly used vehicles).

## Comparing Automated & Human Evaluation Metrics

But can we replace human evaluation with automated evaluation? While we cannot answer this question with quantitative analysis alone, we did two sets of comparisons to get a handle on it:

- Comparing the average human scores to automated semantic similarity to the original definitions to a holistic assessment by the reviewers, given that both attempt to measure the overall correctness of responses.
- Comparing the average human scores to the reading age, given that we also attempted to maximise for being jargon-free and simply written.

To explore these robustly and systematically, we conduct linear and logistic mixed effects models below. To assess whether there is an overall effect of either of the automated scores on the manual scores, we run a linear mixed model with:

- The averaged manual score across reviewers across the four elements (accuracy, absence of jargon, absence of misinformation and appropriateness of example) as the dependent variable
- Automated scores as the independent variable (one model with only one of each of reading age and semantic similarity, and one with both)
- Random effects for terms to account for word-specific variation
- Fixed effects for model type and prompting strategy.

It is worth noting that running a linear model with bounded data, skewed towards the maximum, like this is not ideal, however for interpretability we decided against more complex models. We also are unable to take into account marker effects, due to using the average score between markers.

Overall, we see no meaningful effect of either automated measure on our total human score. This indicates the lack of a strong linear relationship between them. In part, this is because the human scores do account for different concepts to the automated measures, but this still gives an indication that they are both important in considering LLM use cases.

### Table 5: Summary of regression results – total manual marking controlling for automated measures

| Covariate | Average total human score (4-20) | | |
|---|---|---|---|
| **GPT-4** | -0.075<br>(0.072) | -0.059<br>(0.076) | -0.055<br>(0.076) |
| **Zero-shot** | -0.01839 | -0.01795 | -0.01822 |
| **Few-shot** | -0.320***<br>(0.088) | -0.315***<br>(0.088) | -0.316***<br>(0.088) |
| **SentenceTransformers Semantic Similarity** | 0.354<br>(0.336) | NA<br>(NA) | 0.341<br>(0.337) |
| **Dale-Chall Readability Score Change over Reference Definition** | NA<br>(NA) | 0.012<br>(0.015) | 0.011<br>(0.015) |
| **Random effects (% of total variation)** | 21.80% | 21.80% | 21.80% |

**The first five rows give the regression coefficients from the linear mixed effects models, with standard errors in parentheses (* p < 0.05, ** p < 0.01, *** p < 0.001). The final row shows the percentage of total variation explained by differences between terms.**

We then conducted further exploratory analysis to explore relationships with each of the components of the human scores. This time we used the binary scores for each, coded 1 if the average score between markers was 4 or more, 0 otherwise. We then ran logistic mixed effect models with each of these binary variables as the dependent variables, and the models otherwise as stated above. We present these results in Table 6 below.

Here we can see a significant association between semantic similarity and accuracy, with a 0.1 increase in semantic similarity associated with a 1.73 percentage point higher probability of scoring 4 or above in accuracy. Given human markers compared against the original definitions, this alignment is logical. The lack of other significant automated measure effects is notable.

### Table 6: Summary of regression results – breakdown of manual marking controlling for automated measures

| Covariate | Accuracy | Absence of misinformation | Absence of Jargon | Appropriateness of Example |
|---|---|---|---|---|
| **GPT-4** | -0.009<br>(0.012) | -0.008<br>(0.010) | -0.006<br>(0.008) | 0.055**<br>(0.018) |
| **Zero-shot** | -0.042**<br>(0.015) | -0.008<br>(0.011) | -0.024<br>(0.029) | 0.037<br>(0.020) |
| **Few-shot** | -0.008<br>(0.013) | -0.017<br>(0.012) | -0.020<br>(0.024) | -0.037<br>(0.022) |
| **SentenceTransformers Semantic Similarity** | 0.173**<br>(0.055) | 0.031<br>(0.038) | -0.050<br>(0.047) | -0.010<br>(0.062) |
| **Dale-Chall Readability Score Change Over Reference Definition** | 0.004<br>(0.002) | 0.003<br>(0.002) | -0.004<br>(0.004) | 0.004<br>(0.003) |

**The first five rows give the average marginal effects from the logistic mixed effects models, with standard errors in parentheses (* p < 0.05, ** p < 0.01, *** p < 0.001).**

## Error analysis

To better understand cases where automated and human evaluations diverged, we identified distinct patterns of misalignment between our evaluation methods. Our thresholds for identifying misalignments were chosen pragmatically to ensure sufficient cases for analysis whilst capturing meaningful divergence. Specifically, we identified four patterns detailed in Table 7 below.

### Table 7: Error analysis

| Group | Criteria | Number of cases identified |
|---|---|---|
| **1** | High human ratings (≥19 out of 20) but very low semantic similarity[1] scores (bottom 10th percentile) | 15 |
| **2** | Low human ratings (≤16) but relatively high semantic similarity scores (above 70th percentile) | 6 |

[1] As measured by the mpnet-base-v2 measure

| | | |
|---|---|---|
| **3** | High human ratings (≥19) but worsening readability (positive Dale-Chall score change, indicating increased complexity) | 21 |
| **4** | Low human ratings (≤16) but good readability improvements (below 30th percentile of Dale-Chall score changes, indicating substantial simplification) | 7 |

These criteria identified 49 cases of misalignment, with the most common pattern being high human ratings despite poor readability scores (21 cases), followed by high human ratings despite low semantic similarity (15 cases). For qualitative analysis, we randomly sampled five cases from each pattern where available and inspected them. Some high-level patterns from this qualitative analysis include:

**Group 1: High Human Ratings with Low Semantic Similarity (15 cases)**

In these cases, LLM-generated definitions received high human ratings (≥19) despite diverging substantially from the human-written simplified definitions. Terms like 'incentive' and 'inputs' showed good readability while taking different approaches from the reference definitions. Several cases, such as 'services' and 'trade creditors', were noted as being more detailed or covering different aspects of the concept while maintaining clarity. This suggests that LLMs can sometimes produce high-quality simplifications through substantially different approaches than human writers.

**Group 2: Low Human Ratings with High Semantic Similarity (6 cases)**

These cases showed LLM outputs staying close to the human-written simplified definitions yet receiving poor ratings. Despite the high textual similarity, reviewers noted issues with inappropriate or childish examples. Some cases received poor scores across accuracy, jargon, and misinformation, suggesting that close adherence to the reference text didn't guarantee maintaining its quality.

**Group 3: High Human Ratings with Worsening Readability (21 cases)**

These LLM-generated definitions received high human ratings (≥19) despite being less readable than their human-written counterparts. They often used technical terms within the definitions themselves and were generally longer. Some included circular references (such as using jargon to define jargon) yet still received high human ratings. This suggests that human markers valued technical precision above readability.

**Group 4: Low Human Ratings with Good Readability Improvements (7 cases)**

In these cases, LLM outputs achieved better readability scores than the human-written definitions but received low human ratings (≤16). Reviewers consistently noted issues with oversimplification and inappropriate examples. Common criticisms included "childish" or unrealistic examples, with specific concerns about accuracy (rated as low as 2.5) and misinformation. This suggests that while LLMs could sometimes produce more readable text, this didn't necessarily translate to better financial definitions.

# Annex 2: Experiment

## Key Findings

We observed participants' behaviour after receiving LLM-generated savings guidance through a chatbot-style interaction in an online experiment. We explored how it affected participants' abilities to choose the most appropriate cash savings product for the 'profile' they were given, their comprehension of the information they read and their attitudes. We compared this chatbot-style guidance against more traditional Question & Answer (Q&A) content inspired by popular financial guidance websites, as well as a combination of both.

We found that:

- Adding a chatbot onto the Q&A guidance didn't help participants choose the right account.
- Those in the chatbot group were worse at choosing the right savings account compared to the traditional Q&A guidance. Engagement with the chatbot was low, so people saw less information in the chatbot group.
- 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
- Participants' performance in a set of financial comprehension questions was only minimally affected by being in the chatbot group.
- 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 worse performance in choosing the right savings account. This demonstrates that relying on attitudinal survey data may not provide the full picture of consumer-AI interaction, and it is important to have objective measures of behaviour.

## Treatments

Our experiment aimed to compare consumers' responses to financial guidance from LLMs to human-written guidance.

We designed a 'digital assistant' (DA) in the form of a chatbot to provide information to consumers making a choice about which savings account to choose. Faced with constraints around testing consumer interaction with a live LLM, we opted to instead simulate a chatbot experience for participants. The LLM-generated content within the chatbot interface was described to participants as a 'digital assistant' called MoneyChatter. The chatbot was not a live LLM, but gave pre-determined responses to a

selection of questions. Participants could not enter their own questions. This was animated to mimic an online chatbot conversation. The participants were not informed that the content provided by the digital assistant was LLM-generated, nor were they informed that the Q&A information was human-written.

Descriptions of our control group (the Q&A-Only group) and the treatment groups we tested are summarised in Table 1 below. The control group gives a baseline against which we compare the 2 digital assistant treatments, which allows us to measure the effect of these treatments on the outcomes. We created the Q&A-Only group as the baseline as it most closely resembles current practice in guidance. We wanted to understand the effects of adding an option to use a digital assistant to the Q&A information (Q&A+DA), and also fully replacing the Q&A with a digital assistant (DA-Only).

### Table 1: Treatment names, descriptions and characteristics

| Name | Description | Key characteristics |
|---|---|---|
| **Q&A Only (Baseline)** | Financial guidance presented through Question and Answer (Q&A) text. This was inspired by commonly used financial guidance websites, such as MoneyHelper. Information is presented in accordion-style, with question headers as clickable drop-down information. The information is presented using 'chunking', textual priming (question headers), and is not presented all at once to reduce information overload. | • Information provided in Q&A format.<br>• No opportunity to interact with digital assistant.<br>• Information is human written |
| **Q&A+DA** | Presented participants with the same Q&A information as in the control group, but we also offered them to click to launch a digital assistant and 'ask' any of our 3 pre-defined follow up questions, designed to support them in understanding key information. Participants were able to select as many of the pre-defined follow up | • Same information as control group provided in Q&A format.<br>• Opportunity to click to launch the digital assistant<br>• Information in Q&A is human written and digital assistant is LLM-generated |

| | | |
|---|---|---|
| | questions as they liked or to exit. They did not have to interact with the chatbot. | |
| **DA-Only** | We only presented participants with the opportunity to launch the digital assistant, which let them 'ask' any of our 3 pre-defined follow up questions. All of the financial guidance that they received was presented through the chatbot interface and was generated by an LLM after using a different system prompt that asked it to cover particular themes and products. There was no Q&A guidance, so if they chose not to launch the digital assistant, they did not see any information. Participants could select as many of the pre-defined follow up questions as they liked, or they could exit at any time. | • All information provided via the chatbot interface and was LLM-generated.<br>• Opportunity to click to launch digital assistant.<br>• If participants chose not to interact with chatbot, they did not see any financial guidance.<br>• We employed judgement to design effective prompts and to select the best LLM-generated responses. We based this on our hypotheses around the possible benefits of LLM-generated text, such as simplification, brevity, chunking, and reduced search costs by allowing for follow up questions.<br>• We 'engineered' the prompts we gave the LLMs, reviewing the responses the LLMs provided for accuracy and relevance, and adjusting the prompts where needed. |

*The system prompt used was:*

*You are an expert in personal consumer finance, and an experienced financial advisor. Provide simple and easy to understand guidance when asked, as if you were writing to someone with an 8-year old reading level who lives in the United Kingdom. However, please use plain language appropriate to adults. You must not provide financial advice in this instance under any circumstances, complying with Financial Conduct Authority regulations. Prioritise the following when formulating your response: 1) do not give financial advice, 2) ensure the information provided is accurate and relevant and 3) maintain simplicity. When providing guidance avoid technical jargon.*

*Engage the user in a conversational exchange. After imparting clear and concise information, ask open-ended questions that prompt the user to think and respond, facilitating a back-and-forth dialogue. Be sure to complete each segment of information to avoid abrupt endings, and tailor your questions to encourage a natural flow of interaction.*

*Please ensure that each part of your response forms a complete and coherent segment,*

*tailored to the user's query. Be concise but complete within the character limit, avoiding being cut off mid-sentence or mid-thought.*

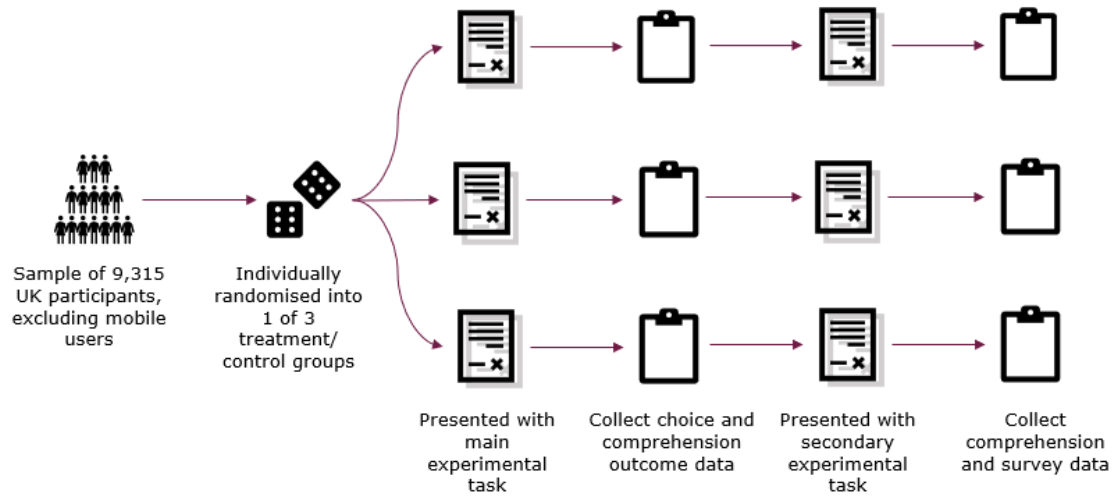
Full details of the input prompts, and of the rest of the text of the experiment are available on request.

## Experimental Design

We conducted an online randomised controlled trial (RCT). Participants were randomly allocated to one of the three treatment or control groups, which means we could infer that any differences in the outcomes that we measured were caused by the treatment we presented them.

Our experiment involved two tasks, which we named the main and secondary tasks (also referred to as Task 1 and Task 2). Participants were provided with the same style of guidance (i.e., LLM-generated Digital Assistant or Q&A, or both) consistently throughout the experiment for both tasks. The flow of the experiment is shown in Figure 1.

### Figure 1. Experimental flow




### Task 1

For Task 1, participants were asked to play the role of a person looking to set up a savings account. They were shown a customer profile, along with 'their situation'.

The situation was that of a confident saver, with emergency savings, who can afford to regularly save £200 a month. This saver did not intend to withdraw their funds for a year and were keen to get the best interest rate. They did not want to invest this money, but to keep it as cash.

The participants were asked to select an account that matched this situation, choosing between an Easy access account, Fixed term savings account, Regular saver account and Cash ISA. They were told that they would see some financial guidance which may help them to find a suitable savings account and help them to learn more about the different features of savings accounts. In the digital assistant treatment groups, they were told

that the information could be reviewed through interacting with the digital assistant. The participants were then taken to a page which showed on of the following guidance types: Q&A-Only, Q&A+DA or DA-Only. We measured participants' engagement with the information provided based on their interaction with the static Q&A sections or the digital assistant.

After reviewing the financial guidance, we measured participants' ability to correctly select the most suitable savings account, which based on, the profile, situation and the account information presented was the Regular Saver. They chose which type of savings account best suited them and their situation and selected a reason for their choice. They then answered 5 comprehension questions about the financial information they had seen.

The initial input prompt for this task for the DA arm was: *My annual income has just gone up to £40,000, and I'm currently setting aside £200 a month into a savings account. Now, I've found I can comfortably save an additional £200 each month. I'm looking for a savings option where this extra money can be put away for at least a year, without the need to access it. Can you outline my options for these additional savings and what factors I should consider in deciding where to place this money?* After the text from this initial prompt was displayed, we allowed a fixed set of follow-up questions in a chatbot-style format.

### Task 2

For the second task, we wanted to evaluate participants' understanding of the information provided. Participants were asked to review some guidance about interest rates (again presented through one of the three guidance variations) and answer 3 comprehension questions (from a multiple choice of 4 options).

We are aware that putting this task second means that participants could have been influenced by having interacted with the digital assistant (or not) in the previous task. To counter this, we could have randomised which task went first out of Task 1 and 2. However, our priority was to generate evidence from Task 1. If we did find such an 'ordering effect', randomising the order would have reduced the amount of data we could analyse from Task 1 without including biased data.

The input prompt for the DA arm for this task was:

*Can you explain how compound interest works?*

### Survey

Finally, we assessed participants' perceptions of digital assistants. Participants filled out an attitudinal survey, where they were asked a series of questions about their attitudes toward digital assistants and asked to write what they would ask a digital assistant if using it for financial information gathering or decision-making. Participants in all treatments were asked these questions, regardless of if they saw or used a digital assistant in Tasks 1 and 2.

# Empirical Strategy

We used regression analysis to estimate the effects of the treatments on our outcomes of interest. These models include covariates for age and gender. In our regression analyses, we corrected for multiple hypotheses testing using the Bonferroni correction approach.

## Primary Analysis

**Performance:** effect of treatment on likelihood of selecting the correct savings account type (Task 1)

**Outcome**: Binary correct (1) or incorrect (0) selection

**Model Specification**: We run an OLS regression using the following model:

$$\mathrm{Y}_i = \beta_0 + \beta_{1-2} X_i + \beta_X X_i + \omega_i$$

Where:

- $Y_i$ is the likelihood of selecting the correct savings account type;
- $\beta_{1-2}$ are the two treatment allocation dummies (one for each treatment group apart from the control);
- $\beta_X$ is the matrix of covariates, as specified below;
- $\omega_i$ are Huber-White robust standard errors.

The matrix of covariates $\beta_X$ includes:

- Gender dummies. Female (base group), Male, Non-binary, Rather not say
- Age group dummies: 18-24 (base group), 25-34, 35-44, 45-54, 55-64, 65-74, 75+, prefer not to say

For **exploratory** analyses, we also ran a separate model to include the covariates $\beta_C$:

- $\beta_C$: Click dummies (binary): Clicked and did not click on chatbot launch button (T1 and T2 only)
- Click interaction term: interaction between treatment and click

## Secondary Analysis

**Comprehension:** Effect of treatment on comprehension

**Outcome**: proportion of correctly answered comprehension questions from 0-1

**Model Specification**: We ran a quasi-binomial regression using the following model:

$$Y_i \sim quasibinomial(n, pi, \phi)\ ;\ logit(p_i) = \alpha + \beta_T T_i + \beta_X X_i$$
$$var(Y_i) = n\phi p_i(1 - p_i)$$

Where:

- $Y_i$ is a two-column integer matrix: the first column gives the number of correctly answered questions (coded 1 for each question), and the second column the number coded as 0 (for each question);
- $T_i$ is a matrix of two treatment allocation dummies (one for each treatment group apart from the control); and
- $X_i$ is the matrix of covariates, as above.

**Outcome: (Exploratory): Effect of treatment on breakdown of comprehension questions**
We ran 5 (Task 1) and 3 (Task 2) OLS regressions with covariates to estimate the impact of treatment assignment on the likelihood of correctly answering each of the comprehension questions.

**Engagement:** effect of treatment on engagement measures

**Outcomes (Exploratory):**

We ran OLS regressions with covariates to estimate the impact of treatment assignment for each of the following:

- Likelihood of clicking digital assistant launch button (Task 1), binary
- Proportion of the 3 follow up questions clicked, given digital assistant was launched (Task 1), from 0 to 1

We also explored the median time spent on the page in seconds (Task 1) for each treatment.

## Sample and attrition

In our study, we collected responses from 9,305 UK adults. We determined our target sample through power analysis, detailed below. Due to technical constraints in coding the experiment, we could only allow participants on desktops or laptops to participate in the experiment. We noted concerns around how the restriction of mobile and tablet users may affect the demographic distribution of our sample, particularly in terms of age. As such, we recruited the sample in batches, allowing for intermittent checks of the demographic data. After inspection of the first 7,700 participants, we determined that the age distribution was indeed becoming more concentrated in the 45-64 age group.

We used a recent, similar large-scale study that we ran as a benchmark for comparison, which allowed for all types of devices to be used. As such, we set out to recruit the remaining participants using quotas on age ranges 18-44, 45-64, and 65+ to mitigate this impact. Whilst we recognise this would still not necessarily result in a fully representative age distribution and may lead to irregularities in the distribution around the cut-off ages, we felt this was an important mitigating step to take to ensure we had adequate samples in our younger and older groups to undertake sub-group analyses and ensure a more age-diverse sample.

The gender identity distribution was 49.9% women, 50% men, and 0.2% as 'prefer not to say'. The median age of participants was 39 years, closely matching the UK's median age of 40.6. Approximately 15% of participants identified as belonging to an ethnic minority background, which is comparable to the 18% of the UK population. Additionally, 48% of participants were in full-time employment, which is lower than the UK's overall employment rate of 75%.

### Attrition

We found that attrition, or those dropping out of the experiment after starting it, was balanced across our treatment groups. Our overall attrition rate was low, with around 1.4% (128 people) dropping out. For the results we report, we included those who dropped out of the experiment if they had been exposed to treatment, coding missing responses as 'wrong' answers. We also ran sensitivity analyses around this approach, such as only analysing complete cases, and found no noteworthy differences. We do not see differential treatment across the arms, seeing 41 non-completes out of 3156 participants in the control group, 35 non-completes out of 3070 participants in Treatment

Arm 1, and 52 non-completes out of 3089 participants in Treatment Arm 2. Chi-squared tests do not find these differences to be statistically significant.

## Approach to missing data

To address missing outcome data, we adopted an intention-to-treat (ITT) approach. Specifically, for participants exposed to treatment, missing or incomplete responses were coded as 0. This method reflected the "lower bound" estimate of the Horowitz-Manski framework, assuming that all missing data corresponded to "incorrect" outcomes.

Missing covariates were handled using the missing indicator method. For both categorical and numeric variables:

- A new indicator variable was created to denote missingness (coded as 1 for missing, 0 otherwise)
- The original variable was coded as 0 wherever it was missing

These missing indicator variables were included in all regression models that used covariates with missing data. This approach ensured that observations with missing covariate values could still contribute to the analysis without biasing parameter estimates.

## Power Analysis

To ensure robust statistical conclusions, we conducted power calculations under the following assumptions:

- Significance level (α): 0.05/2 = 0.025, adjusted for multiple comparisons using Bonferroni correction (2 primary analyses). We correct for multiple comparisons for our primary and secondary analysis only.
- Statistical power: 0.8 (80%)
- Effect size determination: We conservatively chose 50% as our baseline proportion of those correctly choosing the appropriate savings account from our primary outcome measure.

The parameters for the power analysis were therefore:

- Baseline proportion (P1): 0.50
- Minimum Detectable Effect (MDE): 0.0379 or 3.8 percentage points (pp)
- Test type: Two-sided
- Sample size per trial arm (N): 3,300

This sample size was calculated to achieve the stated power and significance thresholds, yielding a total required sample size of 9,900 participants across 3 trial arms. This allocation maximizes power to detect an effect size of 3.8 pp within the constraints of our budget and logistical considerations. The MDE of 3.8 pp was established as a meaningful threshold in consultation with both policy and academic stakeholders.

# Results

### Table 2: Results summary table - comparing the DA treatments against the Q&A-Only baseline.

| | Q&A-Only | Q&A DA | DA-Only |
|---|---|---|---|
| **Performance** | | | |
| **Savings account choice** | 49% | -1pp | **-12pp*** ** |
| **Comprehension** | | | |
| **Average comprehension of savings account information (Task 1)** | 85% | 0pp | **-2pp*** ** |
| Q1: difference between easy access and regular saver | 92% | 0pp | -1pp |
| Q2: penalties for withdrawing early from fixed term account | 94% | -1pp | 0pp |
| Q3: likelihood of earning highest interest rate | 77% | 0pp | **-3pp** ** |
| Q4: effect of interest rate on total savings earned | 88% | +1pp | +1pp |
| Q5: definition of PSA and how it affects taxation of savings interest earned | 75% | 0pp | **-6pp*** ** |
| **Average comprehension of compound interest (Task 2)** | 87% | +1pp | -1pp |
| Q1: effect of compound interest on savings | 90% | +1pp | -1pp |
| Q2: what happens to annual interest rate when on a fixed compound rate | 92% | +1pp | 0pp |
| Q3: how compound interest affects amount of money on which interest is paid | 90% | +1pp | -2pp |
| **Engagement** | | | |

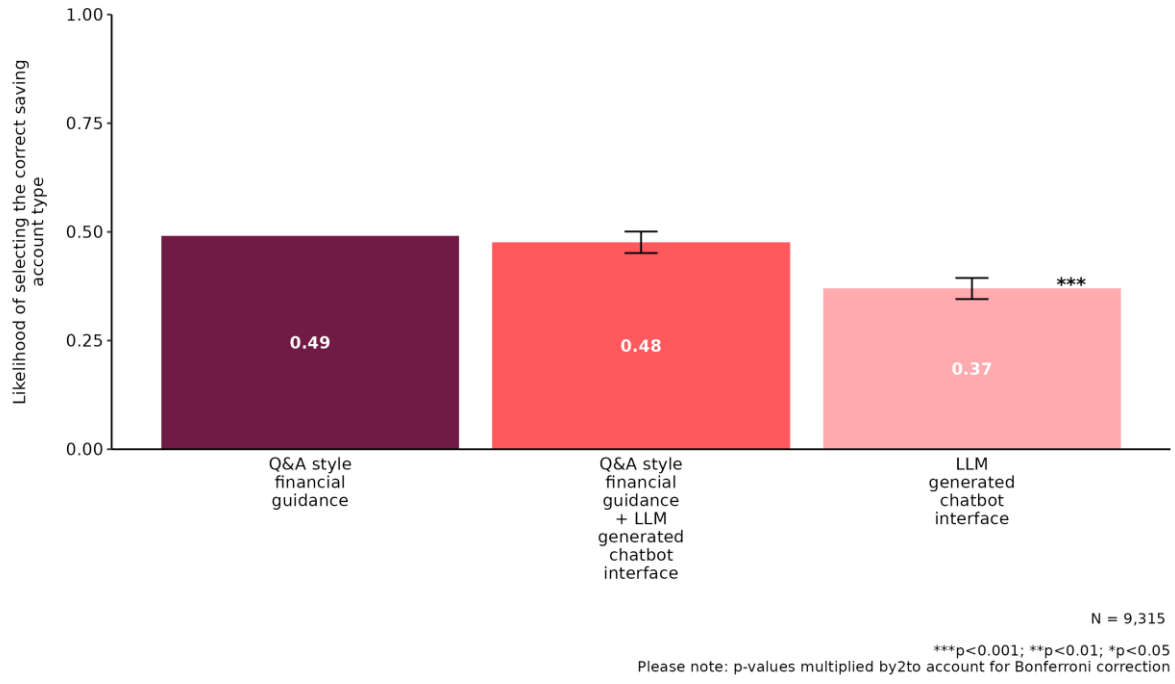
| | | | |
|---|---|---|---|
| Likelihood of clicking digital assistant launch button (Task 1) | N/A | 58% | **+16pp*** ** |
| Proportion of the 3 follow up questions clicked, given digital assistant was launched (Task 1) | N/A | 25% | **-5pp*** ** |
| Median time spent (Task 1) | 131 sec | Didn't launch: 82 sec | Didn't launch: 9 sec |
| | | Launched: 181 sec | Launched: 127 sec |

To note: * indicates significant at the 0.05 level, ** indicates significance at the 0.01 level, and *** indicates significance at the 0.001 level

## Those in the digital assistant-only treatment performed worse at the choosing the appropriate savings account.

Around half (49%) of the participants in the control group identified that the Regular Saver account was the appropriate choice and we saw no difference for the group who were offered the Q&A guidance as well as the digital assistant. Simply adding a chatbot onto existing traditional guidance didn't improve participant's outcomes. However, we found a large, significant effect of -12pp for those who were only offered the digital assistant, pictured below in Figure 2.

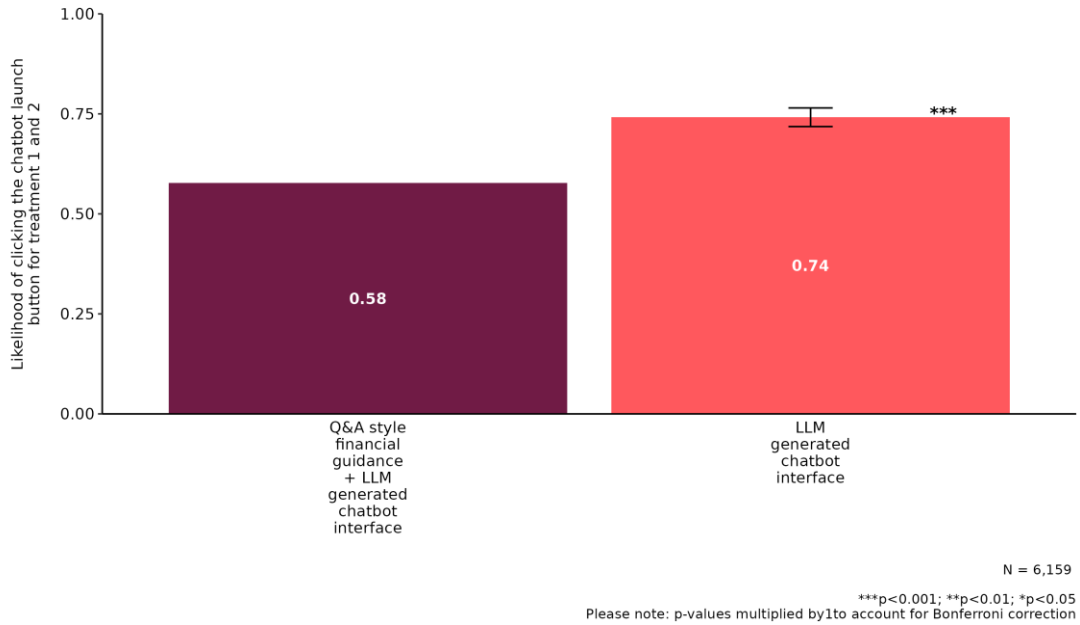
### Figure 2. Effect of treatment on likelihood of identifying the appropriate savings account



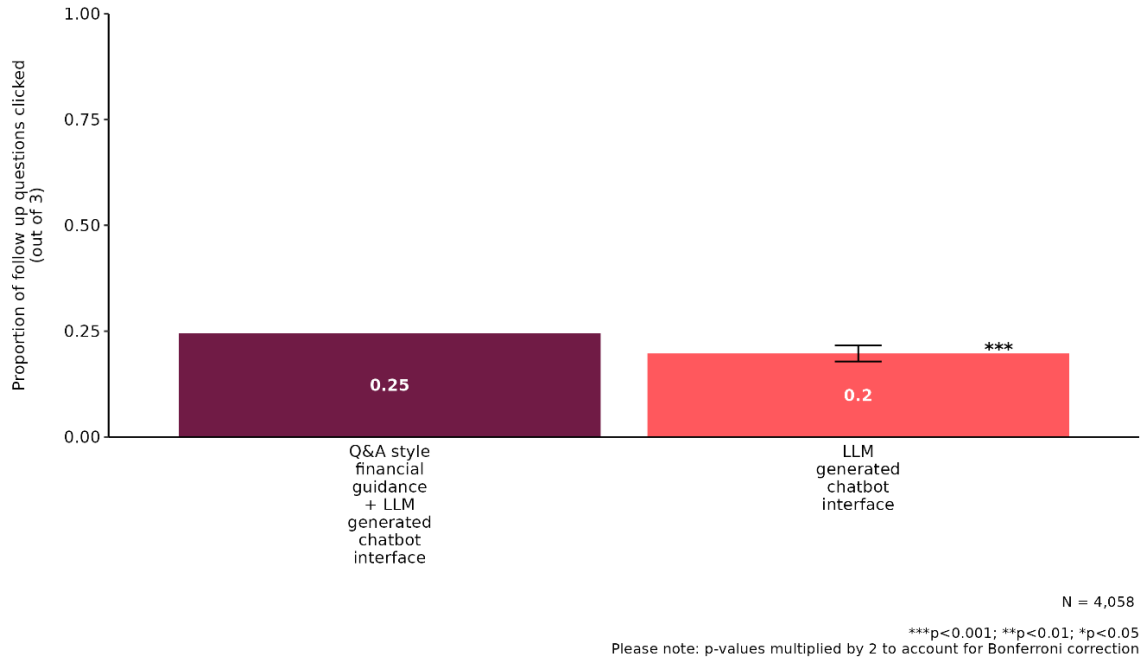
## In general, engagement with the digital assistant was limited.

Around 6 in 10 (58%) participants launched the digital assistant in the group who also saw the Q&A information (Q&A+DA), while around 3 in 4 (74%, +16pp) launched it in the DA-Only group, shown in Figure 3. This means that for those in the DA-Only group, around 1 in 4 did not see any financial guidance at all. For those who did launch the chatbot, engagement with the follow up questions was quite low, with participants clicking around 25% and 20% (-5pp) of the 3 follow up questions, respectively (Figure 4). This means that on average, participants selected less than 1 follow up question.

## Figure 3: Engagement with the digital assistant: proportion who launched



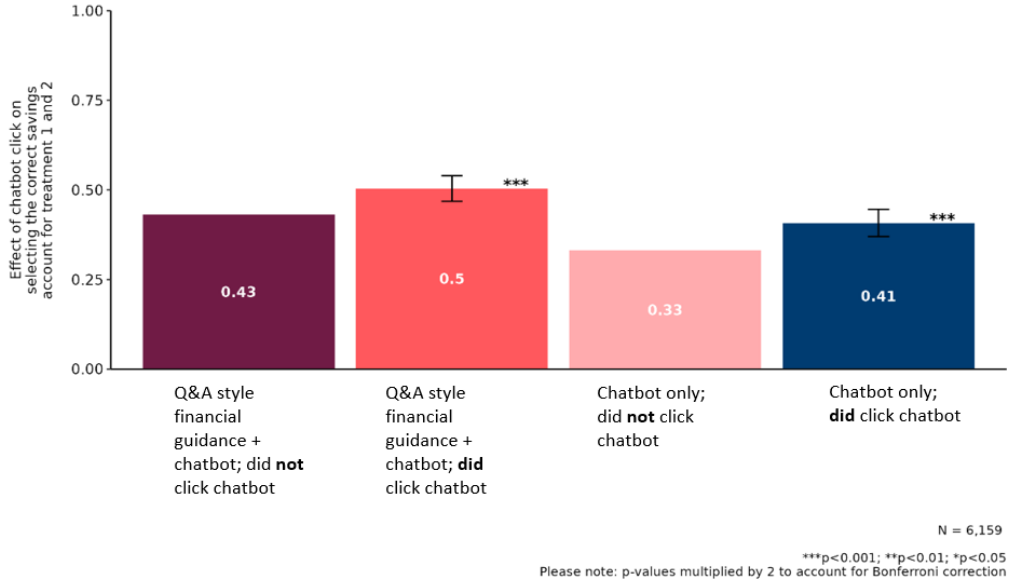
## Figure 4: Engagement with the digital assistant - average engagement with follow up questions.




## Unsurprisingly, higher engagement was associated with better performance.

In both treatments, those who used the digital assistant (by at least clicking to launch it) performed better than those who didn't. While this could be because of a treatment effect, i.e., that engaging more with the information genuinely helped participants in the task, it could also be because of a self-selection effect. In other words, those who engage more with the digital assistant may also be more likely to be interested in this type of information, have more background knowledge, and so on, leading them to perform better.

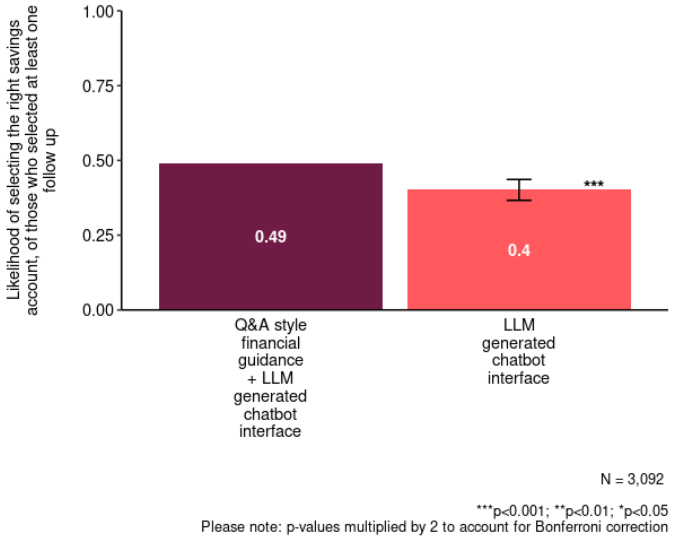
**Figure 5: Comparing those who launched the chatbot to those who didn't – effect on correct choice of savings account.**




## Differences in engagement levels do not explain the negative effect of the digital assistant on choosing the right account.

We initially suspected that differences in participants' abilities to choose the appropriate account may have been driven by their engagement with the digital assistant. As sensitivity analysis, we explored whether it was simply lack of engagement that might have explained the worsened performance in the DA-Only group. This analysis suggested that our main result was not purely driven by lack of engagement, as even the 'highly engaged' in this group performed significantly worse when choosing the appropriate savings account compared to the 'highly engaged' Q&A+DA participants shown in Figure 6. We identified the highly engaged as those who clicked at least 1 of the 3 follow up questions.

## Figure 6. Sensitivity analysis: performance among the 'highly engaged' participants




Though not part of our statistical analysis, we note that the average highly engaged participant in the digital assistant only group performed worse (40%) than the average participant across all engagement levels (i.e., including those who were not highly engaged) in the control group (49%, as seen in Figure 2). This could imply that something about how the information was presented in the digital assistant treatment group led to worse performance on this task.

## Comprehension was much less affected in the digital assistant only group than performance was.

We assessed comprehension in Task 1, around the key information of the savings account products, and Task 2, around compound interest.

For comprehension of Task 1, we found that baseline understanding rates were high, with participants in the control group correctly answering about 85% of the questions. We saw a small, but statistically significant decrease for those in the digital assistant only group of -2pp.We found that this average decrease was driven by comprehension of 2 of the questions, with no statistically significant differences between groups for the other 3 questions.

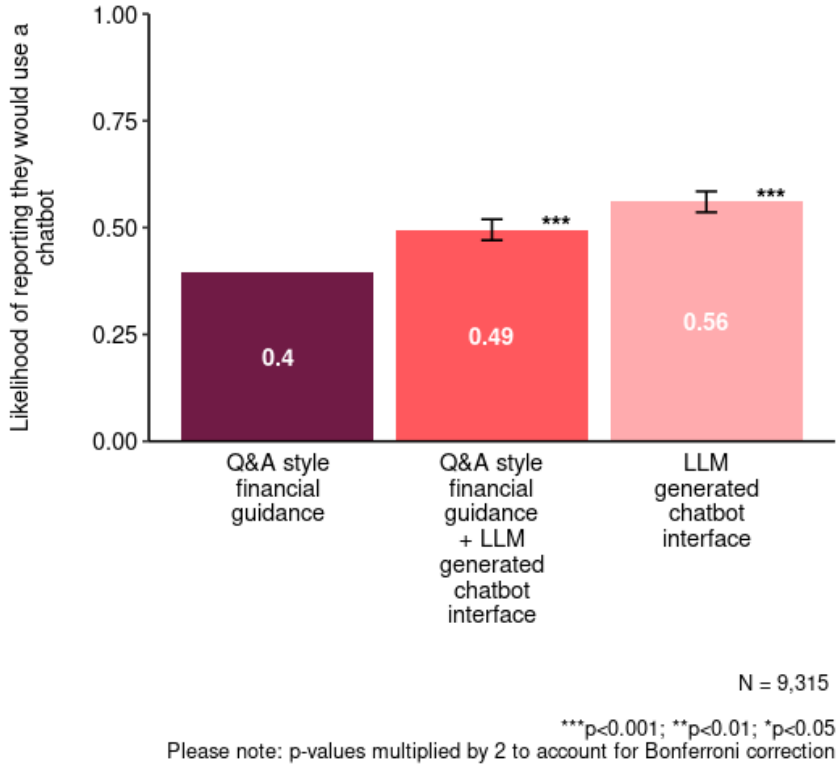
The largest decline of -6pp for the digital assistant only group was for the comprehension question around the taxation of savings interest and the Personal Savings Allowance (PSA). We hypothesise that this may be because of the way that the LLM-generated content explained the PSA as personal savings allowance (i.e., no capitalisation and no acronym) in its first 'message', perhaps making the terminology less recognisable and reducing comprehension.

In Task 2, understanding of compound interest did not differ between groups. Participants demonstrated an overall high level of comprehension in this task, with all groups averaging over 86% correct answers.

## Appetite for the future use of chatbots increased with exposure

The participants who had a chatbot in their guidance were significantly more likely to report that they would consider using a chatbot for financial related decision in the future (Figure 7). This is despite the fact that on average they performed worse at selecting the right savings product.

### Figure 7: Likelihood of reporting they would use a chatbot in the future, controlling for performance on comprehension questions




We note that these attitude questions were asked directly after participants had completed the Task 2 comprehension questions – on which most people performed strongly. This 'recency effect' could have increased confidence in the participants leading to a positive sentiment toward the chatbot. However, when running analysis taking performance in Task 2 into account, we still see a large difference between those in the DA treatments compared to QA only.

## People wanted personalised support to help gather information for financial planning and decision making.

At the end of the experiment participants were asked two questions about chatbots. We conducted exploratory research using an LLM model to synthesise their responses to these free-text questions. This allowed us to analyse a large amount of qualitative data and generate themes and comparisons between groups.

**"How might you use a chatbot to help with making decisions about your finances or to gather information about financial products?"**

Our analysis suggests that participants across all three groups shared common questions on comparing products, understanding financial jargon, and seeking personalised advice. However, each group also showed a distinct focus:

- In the control group, participants primarily asked about strategic advice and savings guidance.
- In the Q&A+DA group, participants were more focused on understanding financial concepts and gathering detailed resources.
- Meanwhile, those in the DA-Only group tended to ask more practical questions, particularly around account selection and making specific product comparisons.

This highlights a key difference: participants exposed to the chatbot tended to prioritize immediate decision-making and resource gathering – similar to the information provided by the chatbot in the experiment - while those in the control group are more inclined to focus on long-term financial planning and strategy. This 'priming' effect likely impacted the ideas participants had.

**"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?"**

Similarly to the first question, participants across all groups sought personalised guidance on savings, investments, and product selection, focusing on strategies to grow wealth and maximize returns. Common 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.

- Participants in the control group showed a specific interest in ISAs and their differences, whereas ISAs were less prominent in the treatment groups. Users in the control group also asked about both short- and long-term strategies, while those in the treatment groups focused more on short-term options.
- Participants in the Q&A+DA group raised broader queries related to financial independence and tax efficiency, as well as questions aimed at understanding basic financial concepts and products.
- Meanwhile, participants in the DA-Only had more practical questions focused on immediate investment gains. Notably, in a sample of responses reviewed, they were more likely to express concerns about the legitimacy of advice and the confidentiality of shared information.

Given these questions were positioned at the end of an experiment concerning savings accounts, the focus on this area is expected. However, participants did show a desire for more detailed information to help with in their financial planning beyond this context and for personalised information.

## Caveats and Limitations

- Due to practical constraints and to give us the ability to compare between participants, the LLM content was created and moderated in advance rather than being a live gen-AI chatbot that participants could interact with freely. There are

many possible outcomes associated with live generated LLM content, and so it is important that this is tested, deployed and monitored carefully.

- We tested only one setting, savings products, where there was a clear correct answer.
- Effects are also likely to be highly context dependent and potentially sensitive to small design choices.
- Comparing Q&A information to chatbot-style content meant that we tested both differences in information presentation and the wording of the information presented simultaneously. This means that we cannot disentangle the effects of each aspect on our outcome variables.
- We speculate that features of how the LLM content was presented could have reduced its impact, such as formatting it as larger chunks of text compared to the Q&A.

# Annex 3: References

Arkoudas, K. (2023). GPT-4 Can't Reason. *arXiv.* Accessed 2nd May 2025 at https://arxiv.org/abs/2308.03762.

Belton, C., Bogiatzis-Gibbons, D.J., Keeley, I., Pi, Y., Spang, J., and Turkay, C. (2025). Credit where credit is due: how can AI's role in credit decisions be explained?. Accessed 2nd May 2025 at https://www.fca.org.uk/publication/research-notes/how-ai-role-credit-decisions-explained.pdf.

Bubeck, S. et al. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv.* Accessed 2nd May 2025 at https://arxiv.org/abs/2303.12712.

Chak, I. et al. (2022). Can robo-advice improve borrower payment decisions? *FCA Occasional Paper 61.* Accessed 2nd May 2025 at https://www.fca.org.uk/publications/research/can-robo-advice-improve-borrower-repayment-decisions.

Chernev, A., Böckenholt, U., and Goodman, J. (2015). Choice overload: A conceptual review and meta-analysis, *Journal of Consumer Psychology*, 25(2): 333-358. https://doi.org/10.1016/j.jcps.2014.08.002.

Dell'Acqua, F. et al. (2023). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper No. 24-013, The Wharton School Research Papers*. Accessed 5th May 2025 at: https://ssrn.com/abstract=4573321.

Deng, N., Hodroj, B., Latham, A. J., Lee-Tory, J., & Miller, K. (2024). Is present-bias a distinctive psychological kind? *Inquiry*, 1–27. https://doi.org/10.1080/0020174X.2024.2321614.

Dwyer, L., Francis, W. and Tyagi, S. (2025). A Pilot Study into Bias in Natural Language Processing. Accessed 2nd May 2025 at https://www.fca.org.uk/publication/research-notes/pilot-study-bias-natural-language-processing.pdf.

Eidelman, S. and Crandall, C.S. (2012). Bias in Favor of the Status Quo. *Social and Personality Psychology Compass*, 6: 270-281. https://doi.org/10.1111/j.1751-9004.2012.00427.x.

Federal Reserve Bank of St. Louis. (2025). Glossary of Economics and Personal Finance Terms. Accessed 2nd May 2025 at https://www.stlouisfed.org/education/glossary.

Fieberg, C., Hornuf, L., Meiler, M., and Streich, D. (2025). Using Large Language Models for Financial Advice. *CESifo Working Paper No. 11666*. Accessed 2nd May 2025 at http://dx.doi.org/10.2139/ssrn.5133294.

Gu, J. et al. (2024). A Survey on LLM-as-a-judge. *arXiv*. Accessed on 2nd May 2025 at https://arxiv.org/abs/2411.15594.

Hsiao, Y-C., Kemp, S., Servátka, M., Ward, M., and Zhang, L. (2021). Time Costs and Search Behavior. *Munich Personal RePEc Archive.* Accessed on 2nd May 2025 at https://mpra.ub.uni-muenchen.de/105412/1/MPRA_paper_105412.pdf.

Hugging Face. (2024). sentence-transformers/all-mpnet-base-v2. Accessed on 6th May 2025 at https://huggingface.co/sentence-transformers/all-mpnet-base-v2.

Ichien, N., Stamenković, D., and Holyoak, K.J. (2024). Large Language Model Displays Emergent Ability to Interpret Novel Literary Metaphors. *arXiv*. Accessed on 2nd May 2025 at https://arxiv.org/abs/2308.01497.

Meyer, J. et al. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence*, 6. https://doi.org/10.1016/j.caeai.2023.100199.

Money & Pensions Service. (2021). 24 million UK adults don't feel confident managing their money, Talk Money Week is here to help. Accessed on 2nd May 2025 at https://maps.org.uk/en/media-centre/press-releases/2021/24-million-uk-adults-dont-feel-confident-managing-their-money-talk-money-week-is-here-to-help#.

O'Donoghue, T., & Rabin, M. (1999). Doing it now or later. *American Economic review*, *89*(1), 103-124. https://www.aeaweb.org/articles?id=10.1257/aer.89.1.103

OpenAI et al. (2023). GPT-4 Technical Report. *arXiv.* Accessed on 2nd May 2025 at https://arxiv.org/abs/2303.08774.

OpenAI (2024). GPT-4 Model Card. Accessed on 6th May 2025 at https://cdn.openai.com/papers/gpt-4-system-card.pdf.

Plain English Campaign. (2023). The A to Z of financial terms. Accessed on 2nd May 2025 at https://web.archive.org/web/20231204030130/https:/www.plainenglish.co.uk/files/financialguide.pdf.

Ross, J.A. (2024). Examining LLMs in Economic Settings (MSc Thesis). Accessed on 2nd May 2025 at https://dspace.mit.edu/bitstream/handle/1721.1/156339/ross-jillianr-sm-eecs-2024-thesis.pdf?sequence=1&isAllowed=y.

Weidinger, L. et al. (2022). Taxonomy of Risks posed by Language Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 214–229. https://doi.org/10.1145/3531146.3533088.

White, J. et al. (2023). A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. *arXiv*. Accessed on 6th May 2025 at https://arxiv.org/abs/2302.11382.

Zhou, Y et al. (2024). Are LLMs Rational Investors? A Study on Detecting and Reducing the Financial Bias in LLMs. *arXiv*. Accessed on 2nd May 2025 at https://arxiv.org/pdf/2402.12713.

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