Content Analysis of LLM Financial Advice Responses

Jack Marley-Payne

2026-03-19

Abstract

This study presents a preliminary content analysis of responses generated by four leading large language models (LLMs) — ChatGPT, Claude, Grok, and Gemini — answering common personal finance questions. Our focus is providing an assessment of the responses in terms of the quality of the *financial advice* provided. Using an automated classification pipeline built around the Anthropic API, we coded all statements in each response by communicative type and applied sub-analyses including fact-checking, sentiment analysis and jargon counting. We found that all models provided a high volume of advisory and factual statements, skewed confident in tone, and included high levels of jargon. They made limited efforts to seek the additional context that best practices in financial advising would recommend acquiring before distributing advice. Factual accuracy was generally high, with errors being the exception not the rule. There were, however, significant differences between models. Notably, Grok and Gemini produced higher rates of vague or misleading factual claims, while ChatGPT produced a disproportionate number of leading questions. Claude tended to be relatively brief and cautious in its responses. These findings raise questions about the quality and consistency of financial guidance that general users are likely to receive from popular LLMs, and motivate more systematic investigation of this rapidly evolving domain.

Introduction

Use of large language models (LLMs) by both consumers and professionals has spread rapidly in recent years, including in relation to financial decision making. One consequence of this is that many people are now seeking guidance on personal finance through a new medium for which much is still unknown.

Unlike traditional financial advice given by a financial adviser or financial counselor — which is regulated, credentialed, and subject to fiduciary obligations — advice generated by general-purpose AI chatbots lacks regulatory safeguards. In contrast with online searches on financial topics, though, which simply present a list of websites with information related to the search text, LLMs provide a fluent conversational tone that engages with the user directly, which may give the 'feel' of speaking to a financial adviser. Further, the confidence with which LLMs present information may lead users to place undue trust in their outputs. On the other hand, these tools have the potential to provide information on demand, synthesized from multiple sources and tailored to specific consumer needs, in a way many might not otherwise be able to access.

This paper aims to expand our knowledge of this pressing issue. We conduct a preliminary study of what four leading LLMs actually say when asked common personal finance questions, using a systematic content analysis framework. We see that LLM use provides both potential benefits and risks for consumer finance, as well as potential guidance on how they might be used most effectively.

Literature Review

Use of LLMs has expanded rapidly in recent years (Chatterji et al., 2025). This includes a significant number of people asking these tools for financial advice: according to a recent survey, representative of the population, around 20% of US adults are interested in receiving financial advice from AI (FINRA, 2025). This has the potential to significantly change people's financial behavior.

However, relatively little is known about how LLMs are being used for personal financial advice specifically. Most extant studies have focused on professional and institutional applications rather than on the advice that ordinary consumers receive from general-purpose chatbots (Li et al., 2024; Yie et al., 2024). Where consumer-facing use has been studied, research has tended to measure whether particular populations use or wish to use LLMs to assist their personal finances (FINRA, 2024), or whether the information on financial topics such as mortgages is accurate (Galadima et al., 2025). An additional concern is evaluating LLMs as *advisers*, which requires thinking about how average users interact with AI and analyzing the responses systematically from multiple angles.

Many concerns around LLM use are particularly acute in personal financial settings: namely that such tools provide inaccurate or misleading information, especially through “hallucinations” — instances where a model generates confident but incorrect or fabricated information (Huang et al., 2025) — or that they encourage over-confidence in users (Kim et al., 2024). These worries may well also influence consumer behavior given the widespread attention paid to them in popular media (BBC/EBU, 2025; BBC News, 2026).

The concern about over-confidence is especially salient given that there is a well-established body of evidence showing that financial over-confidence leads to negative real-world consequences, including excessive trading, higher transaction costs, increased risk-taking, and higher rates of mortgage delinquency (Inghelbrecht & Tedde, 2024; Kim et al., 2020). On the other hand, LLMs have the potential to be a powerful source of financial knowledge, and research has shown that higher financial knowledge leads to better financial outcomes (Lusardi & Mitchell, 2014).

Finding out and analyzing the financial advice LLMs provide as of today is a pressing issue, especially as updates to the models dramatically change the nature of responses being offered on a yearly or even monthly basis (Galadima et al., 2025).

Content Analysis of AI-Generated Text

Content analysis is an established method for systematically examining the content of communication (Krippendorff, 2022). The approach has recently been extended to encompass automated and semi-automated coding using large language models. Chew et al. (2023) propose an LLM-assisted content analysis framework and find that it can often perform deductive coding at levels of agreement comparable to human coders when given a well-specified codebook. Tai et al. (2024) similarly find that LLMs produce consistent and reliable results that qualitative researchers can use to validate traditional coding. The use of an LLM as classifier in the present study, where the *subject* of analysis is itself LLM-generated text raises some particularly salient questions regarding methodology — an issue we return to in the Discussion.

To our knowledge, no prior study has applied systematic content analysis to compare the rhetorical and informational characteristics of financial advice responses across multiple leading LLMs. Prior work on AI financial advice has largely focused on consumer attitudes and accuracy benchmarking (FINRA, 2024; Galadima et al., 2025) rather than on evaluating the quality and impact of responses as financial *advice* with an eye on how users may respond to receiving such guidance.

Methods

This study analyses the responses of four popular LLMs — ChatGPT, Claude, Grok, and Gemini — to common personal finance questions. The goal is to understand what kind of advice typical users are likely to receive from such models. As data on the precise questions being asked and responses given is not currently available, we approximated them based on common financial questions asked in general.

We looked at data on the most commonly searched personal finance questions — through a search engine, not LLMs — (The Currency, 2024; Liberty Bank, 2019) and cross-referenced these with financial education resources on the key areas of personal finance (CFPB, 2024; OECD/INFE, 2020). This gave us the following five questions:

1. How do I invest in stocks?
2. How can I save money?
3. What are the benefits of budgeting?
4. How do I make sure I have enough money for retirement?
5. How can I pay off my debt?

In addition, we asked five alternative questions on each topic, to assess how sensitive responses were to specific wording:

1. Should I invest in stocks?
2. What are ways to save money?
3. What are the pros and cons of a budget?
4. How do you plan financially for retirement?
5. What are approaches to dealing with debt?

In our judgment, though these questions differed somewhat in their literal meaning, casual LLM users with general questions about each topic might well not distinguish them and could easily ask either one while having the same needs. Therefore, it would not be serving these users well to provide radically different responses depending on the version of the question being asked.

We gave these questions to four of the leading LLMs for everyday users: ChatGPT, Claude, Grok, and Gemini.

As we are trying to approximate an unsophisticated user of these tools, we input each question into a new chat with no additional context or instructions. We used the default free version of each model using an account with no previous history. The models used were as follows: ChatGPT - GPT 5.2; Claude - Sonnet 4.5; Grok - Grok 4.1; Gemini - Gemini 3. We asked all questions between 2/4/2026 and 2/5/2026.

We took these responses and analysed them using both qualitative and quantitative content analysis.

After informal inspection of the raw data, and given the concerns surrounding LLM use outlined above, we arrived at the following analytical goals:

a. Identify any factual claims made and check whether they were accurate.
b. Identify advice given and assess how forceful the recommendations were.
c. Examine whether questions sought to clarify more about the user's context or to nudge them towards particular conclusions.
d. Assess whether responses trended towards cautious or confidence-boosting framing, and how much financial jargon they contained.

This led us to the following analysis plan.

Qualitative Content Analysis

The LLM responses (stored as text files) were processed through a multi-step automated pipeline built around the Anthropic API, which used the *Claude* LLM. This pipeline was designed to prompt Claude to first evaluate each statement in the LLM response and decide which of the following six categories it belonged to: factual, advice, normative, question, self-referential, and miscellaneous. Second, depending on the category assigned, it had to assign the statement to the appropriate sub-category:

- **Factual**: precision (precise/vague)
- **Advice**: strength (suggestion/recommendation/imperative)
- **Questions**: function (clarifying/leading/rhetorical/reflective)
- **Self-referential**: direction (confidence/capability/limitation)

A full technical explanation of the pipeline is provided in the appendix. The coding scheme was developed inductively through iterative testing on sample chunks from the data, with the prompt refined in response to observed misclassifications before being applied to the full dataset. This inductive development approach follows recommended practice for LLM-assisted content analysis (Chew et al., 2023).

All LLM responses were stripped of identifying information as to the model used before being submitted for analysis.

Fact-Checking

Factual statements identified in the classification stage were extracted from the main dataset. These statements were then submitted to Claude via the Anthropic API for verification, instructing the model to assess each claim and assign one of five labels: Unambiguously true, Vague but generally true, Vague but misleading, Unambiguously false, or Unable to evaluate.

Quantitative Sentiment and Jargon Analysis

A separate LLM-assisted pass over the full response texts counted occurrences of positive phrases, negative phrases, and financial jargon per response. This was conducted at the response level rather than statement level and merged with the main dataset.

Cross-Wording Consistency Analysis

To assess sensitivity to question framing, statements from the primary question wording were compared to their counterparts in the alternative wording using Claude as a classifier. Each statement from version 1 was assigned one of three match labels relative to version 2: complete match, partial match, or no match.

Reliability

To assess the reliability of the LLM assisted coding, results were compared with those produced by a human coder using a codebook. Reliability was assessed at two levels using two separate samples.

For primary label reliability (Sample A), a stratified random sample of 100 statements was drawn from the full corpus, with sampling proportional to model, question topic (Q1–Q5), and statement type. An independent human coder, working from a purpose-built codebook, assigned a primary label (factual, advice, normative, question, self-referential, or miscellaneous) to each sampled statement without access to the automated classifications. Primary label agreement was then evaluated across all 100 statements using Krippendorff's alpha with a nominal metric, the standard measure for categorical content analysis data (Krippendorff, 2022; Hayes & Krippendorff, 2007).

For key signal reliability (Sample B), a separate stratified random sample of 30 statements per type was drawn for each of the four types carrying key signals (factual, advice, questions, and self-referential), yielding 120 statements in total. The human coder was informed of the statement type for each item in Sample B. The coder then assigned the appropriate key signal without access to the automated classifications. Key signal alpha was reported both pooled across all four types and separately by category.

The sentiment and jargon analysis operates differently from the classification pipeline: rather than assigning a categorical label to each statement, it identifies and extracts specific phrases within statement text. Standard inter-rater reliability metrics such as Krippendorff's alpha are not appropriate for this task. Reliability was therefore assessed using precision, recall, and F1, which separately quantify over-flagging (low precision) and under-flagging (low recall) relative to human annotation.

A stratified random sample of 60 statements was drawn from the full corpus (15 per model), with at least 60% of each model's allocation drawn from statements containing at least one automated phrase flag, to ensure sufficient phrase instances for stable estimates. A human annotator, working without access to the automated output listed all positive, negative, and jargon phrases present in each statement.

Because fact-checking is fundamentally a validity question rather than an inter-rater reliability question — two reviewers could independently agree on a label and both be wrong — a standard alpha-based reliability analysis is not the appropriate check. Instead, a human auditor independently verified a sample of 30 statements assigned one of the two unambiguous labels (Unambiguously true or Unambiguously false) against authoritative sources (e.g., IRS publications, Federal Reserve data). These are the labels where the pipeline expressed the strongest certainty, and therefore where label errors would be most consequential. Statements assigned vague or "unable to evaluate" labels were not subject to audit. Undertaking analysis that evaluated the LLM driven fact checking more systematically was beyond the scope of this preliminary analysis but would be a valuable project for further research.

Reliability Results

An alpha value of ≥ 0.80 is generally considered satisfactory for drawing reliable conclusions, with values between 0.67 and 0.79 considered sufficient for tentative conclusions (Krippendorff, 2022). Primary label reliability (Sample A, n = 100) yielded a Krippendorff's alpha of 0.83, exceeding the 0.80 threshold. Key signal reliability (Sample B, n = 120) yielded a pooled alpha of 0.823. Results varied substantially by type, however, and should be interpreted at the type level rather than from the pooled figure alone.

Question function classification achieved an alpha of 0.754, sufficient for tentative conclusions, with high agreement on clarifying (100%) and leading (89.5%) labels. The reflective label showed lower agreement (33.3%), though the small number of reflective statements in the sample (n = 3) limits interpretation.

Self-referential direction achieved an alpha of 0.802, satisfactory for drawing conclusions. Agreement was high for confidence (95.8%) and limitation (100%) labels. The capability label was assigned only once in the sample, making per-label estimates unreliable for that value.

Factual precision achieved an alpha of 0.655, falling below the threshold for reliable conclusions. This figure should be interpreted cautiously: 28 of 30 statements were classified as vague by the automated classifier, leaving only 2 precise statements in the sample. The near-zero variance on the precise label inflates apparent disagreement in the alpha calculation. Findings relating to factual precision should nonetheless be treated as tentative.

Advice strength achieved an alpha of 0.371, indicating poor agreement between the automated classifier and the human coder. Agreement was below 70% across all three signal values (imperative: 69.2%; recommendation: 57.1%; suggestion: 66.7%), suggesting that the boundary between these categories — particularly between imperative and recommendation — presents interpretive difficulty that the codebook does not fully resolve. Results relating to advice strength should be interpreted with caution, and differences between models on this dimension treated as indicative only.

Phrase detection reliability was acceptable across all three categories (Table R2). Jargon detection performed best (F1 = 0.833), with equal precision and recall. Positive phrase detection also performed well (F1 = 0.786), again with balanced precision and recall. Negative phrase detection was slightly lower (F1 = 0.718), with a modest precision shortfall (0.700 versus recall 0.737). Pooled across all three categories, F1 was 0.769, indicating broadly adequate detection reliability.

All fact-checking statements validated by a human auditor were confirmed correct. This was an informal check, however, and fact-checking results should be treated accordingly.

Results

The tables below report results across the four models. Note that raw statement counts are not directly comparable across models because response length varies substantially — Grok produced approximately twice as many total words as Claude (see Table 6). Proportional comparisons within each table may at times be more appropriate than absolute counts.

Table 1. Statement type counts by model

Model	Advice	Factual	Miscellaneous	Normative	Question	Self Referential
ChatGPT	163	115	19	83	32	24
Claude	102	77	15	27	13	11
Gemini	125	178	21	67	17	1
Grok	201	210	18	65	19	14

Grok produces the largest total number of statements across all types. ChatGPT is notable for a relatively high count of self-referential statements compared to other models, and for producing more than twice as many questions as any other. Gemini produces only a single self-referential statement across all ten questions. Claude has the lowest number of statements overall and indeed the lowest value for each category beside self-referential statements.

Table 2. Advice strength by model

Model	Imperative	Recommendation	Suggestion
ChatGPT	61	82	20
Claude	35	51	16
Gemini	53	59	13
Grok	67	109	25

All four models produce more recommendations than either imperatives or suggestions, though the balance differs. ChatGPT has the highest proportion of imperatives relative to its total advice count. Claude has the most even distribution across the three strength levels.

Table 3. Question function by model

Model	Clarifying	Leading	Reflective	Rhetorical
ChatGPT	7	22	3	0
Claude	11	1	1	0
Gemini	6	8	2	1
Grok	6	13	0	0

The most striking pattern here is ChatGPT's use of leading questions (n = 22), compared to between 1 and 13 for the other models. Claude produces far more clarifying questions (n = 11) than any other model, suggesting a different orientation towards the user.

Table 4. Self-referential statement direction by model

Model	Confidence	Capability	Limitation
ChatGPT	24	0	0
Claude	3	1	7
Gemini	1	0	0
Grok	11	1	2

ChatGPT’s self-referential statements are exclusively confidence-oriented — the model never acknowledges a limitation or offers a capability claim. Claude is the only model to produce a substantial number of limitation statements (n = 7), suggesting meaningfully different epistemic stances across models.

Table 5. Fact-check labels by model

Model	Unable To Evaluate	Unambiguously True	Vague But Generally True	Vague But Misleading	Unambiguously False
ChatGPT	11	43	60	1	0
Claude	4	46	26	0	1
Gemini	11	81	73	9	4
Grok	16	101	81	7	5

Grok and Gemini produce notably higher counts of vague but misleading and unambiguously false factual statements (12-13 each). ChatGPT and Claude each only produced one false or misleading statement.

Table 6. Sentiment and jargon counts by model

Model	Positive Phrases	Negative Phrases	Jargon Terms	Total Words
ChatGPT	193	127	120	3,770
Claude	60	45	80	2,459
Gemini	133	96	156	4,362
Grok	197	104	257	4,952

Grok is the most verbose (4,952 total words) and uses substantially more financial jargon (n = 257) than the other models. Claude is the most concise (2,459 words) and uses the least jargon (n = 80). Positive framing dominates over negative framing across all models.

Table 7. Cross-wording consistency by model

Model	Complete Match	No Match	Partial Match	Match %
ChatGPT	104	54	54	74.5%
Claude	56	38	33	70.1%
Gemini	80	76	33	59.8%
Grok	91	101	43	57.0%

Match percentages range from 57.0% (Grok) to 74.5% (ChatGPT), indicating that between a quarter and almost half of statements do not have a counterpart under alternative question wording. ChatGPT and Claude show greater consistency than Grok and Gemini, suggesting their responses are less sensitive to surface-level variation in question phrasing.

Discussion

The results above reveal both common trends along with substantial heterogeneity in how the four models approach personal finance questions — not only in terms of length and content, but also in rhetorical tone. Several patterns are worth highlighting. Across all models, the density of information in each response was significant, providing on average between 10-20 facts and 10-20 pieces of advice per question. Though Claude was notably lower in quantity than other models, it was still high relative to what one might hope for with financial advice, where best practices suggest gaining additional context about a consumer's needs and background before offering detailed advice (Theodos et al., 2015). Relatedly, the number of questions asked by the LLMs was much lower than advice and facts, between 1 and 3 per response.

Users faced with this volume of information may either be overwhelmed and disengage, ignore it and just focus on the “next steps” provided in the final paragraph, or skim it and increase their confidence without having fully understood what they take themselves to have read.

Advice style and directiveness. All four models are predominantly advisory in character. The ratios between type of advice were relatively similar across models with recommendations the most common, then imperatives, then suggestions. This tendency towards issuing stronger advice without gathering much context about the user is potentially concerning and is in general against best practices in financial counselling (Theodos et al., 2015). Another problem is the sheer quantity of advice, 10-20 items per question, which is unlikely to all be followed unless it’s incorporated within a structured plan of action.

This study does not evaluate the quality of the advice being provided, however, which is an important topic for further investigation. From observation, much of the advice corresponds to well-known rules of thumb in the personal finance world, such as recommending the “50/30/20” budgeting rule. Though not terrible advice in a vacuum, in general one would want to gather more information about a consumer before deciding if such a strategy was appropriate. At times, the advice could be extreme, such as Grok instructing the user to always put their entire tax refund and work bonus into savings — which might be counter-productive for someone having difficulties keeping up with their daily expenses.

Questions. One area of note is the distribution of question types: all models aside from Claude ask more leading questions than clarifying questions despite having no background information about the user. ChatGPT’s heavy use of leading questions — more than all other models combined — is particularly striking. Leading questions function rhetorically to nudge users towards conclusions without making those conclusions explicit; in this case, these generally involve nudging them towards engaging in a specific plan of action. For example, one model's response to a budgeting question asked whether the user would like help drawing up a "50/30/20" budget, rather than seeking the context to decide whether such an approach was best for them. This aligns with worries about LLMs pushing users into taking drastic actions without fully thinking them through.

Also noteworthy is the numerical lack of clarifying questions: only Claude asks more than one per query on average. This is a concern given that the appropriate ways to tackle the broad financial issues raised in these questions are highly dependent on individual circumstances.

Self-referential framing.

There is significant heterogeneity with regard to the models' self-referential statements. Only Claude provides more expressions of limitation than confidence. In particular, Claude notes multiple times that it is not a financial advisor and recommends seeking out advice from such a source. ChatGPT's self-referential statements, on the other hand, are exclusively confidence-oriented. Grok is similarly skewed towards confidence while Gemini offers barely any self-referential statements of any form.

A model that never acknowledges uncertainty or limitation may be more persuasive in the short term, and may be more enjoyable for the user to engage with. Indeed, in promoting a recent update, OpenAI reports that its new model offers fewer "disclaimers and caveats" in its responses, stating that this is what users are asking for (OpenAI, 2026). However, this tone may also contribute to the problematic over-confidence effects in consumers documented in the personal finance literature (Inghelbrecht & Tedde, 2024; Kim et al., 2020). The concern raised by

Huang et al. (2025) — that LLMs present information with high fluency and confidence even when its appropriateness is uncertain — appears directly relevant here. One of the standards listed in the Association of Financial Counselors code of ethics is “Recognize my limitations and refer clients when appropriate.” (AFCPE, 2018)

Factual accuracy and precision. The results regarding factual accuracy are nuanced. Statements are for the most part accurate, running counter to the most extreme worries about widespread LLM inaccuracies. Indeed, both ChatGPT and Claude only have one false or misleading statement across all questions. The higher rates of vague but misleading and unambiguously false statements for Grok and Gemini do raise concerns about LLM hallucination in financial contexts (Galadima et al., 2025; Huang et al., 2025). That said, most of these falsehoods, when examined, involve relatively minor points, such as confusing the 2024 and 2025 IRA contribution limits, stating it was $23,000 not $23,500. Though all inaccuracies are a cause for concern, these are not the kind of mistakes likely to lead someone to financial catastrophe, unlike some of the more notorious examples of LLM hallucination in other areas.

That a non-trivial proportion of factual claims across all models fall into the “vague but generally true” category also warrants attention: vague claims may be harder to falsify, but they may also be less useful to users trying to make concrete decisions. Many of these statements seem more like conversational rapport building than presentation of pertinent information - e.g., “budgeting gets a boring reputation.” Though unlikely to directly mislead, these may work to build trust and give the user the feeling of talking to a personal advisor, which as mentioned above can be a double-edged sword.

Outside the scope of the analysis were statements that were unambiguously true, but could lead to potentially misleading inferences. For example, in response to the question “Should

I invest in stocks”, Grok included the statement, “Goldman Sachs projected ~12% total return for the S&P 500 in 2026”. While this statement is correct, crucial omitted context is that such start of year projections have historically had little predictive value. This points to a broader potential problem as to whether current LLMs have sufficient holistic reasoning capabilities to avoid misleading users.

Sentiment and Jargon Analysis In line with expectations, all responses skewed heavily towards confident phrases over cautious ones. The potential risks here have already been discussed. Also of note are the high levels of jargon included in responses, especially for Grok which averaged nearly 26 pieces of jargon per response.

The effects of subjecting users to so much unfamiliar vocabulary are unclear. It could lead to them feeling overwhelmed and disengaging from the conversation before being able to make an informed decision. Reading through a jargon-filled block of text could also make someone feel like an expert even if they didn’t truly understand what was being said, leading to potentially problematic overconfidence. On the other hand, it could cause them to look up definitions of unfamiliar terms and so expand their knowledge.

Based on our observational analysis of the responses, one reason for concern is that many of the models introduced jargon loosely in contexts that could be misleading. For example, Grok, in the same response as before referred to “index funds/ETFs” as a way to acquire a diversified investment. This could lead someone to infer the two are interchangeable. However, though many ETFs do track broad index funds while offering low fees, other ETFs are extremely risky, often designed to provide greater volatility than individual stocks.

Cross-wording consistency. The relatively low match rates — particularly for Grok and Gemini — suggest that users asking substantively identical questions in different words may receive materially different advice. This is a neglected dimension of LLM reliability: even if a model performs well on average, high sensitivity to phrasing means that user outcomes will depend substantially on how they happen to phrase their question.

Much of this reflects literal question interpretation. For example, if you ask, “What are the benefits of budgeting?” you are not literally asking about any potential downsides while if you ask, “What are the pros and cons of budgeting?” you are. Accordingly, many of the models only listed such downsides for the latter question. However, it is plausible that many users will not be thinking so carefully about the specific presuppositions of the wording of their prompt. They might, for example, be deciding whether to create a budget when they ask “What are the benefits of budgeting?”, and so the additional context about drawbacks would in fact be relevant for them.

Limitations

Use of an LLM as classifier. A methodological limitation of this study is that the classification pipeline itself is built around an LLM, meaning that the analysis of LLM-generated text is conducted by another LLM. While Chew et al. (2023) and Tai et al. (2024) find that LLM-based deductive coding can match human performance in general, the use of LLMs to analyze LLM response text appears a special case. One might even be concerned that LLMs are instructed to cast themselves in a positive light. Though, to be clear, no current evidence has come to light suggesting this is the case.

The inter-coder reliability analysis provides partial evidence on this question. Primary label agreement was satisfactory ($\alpha = 0.83$), and key signal reliability was adequate for question function ($\alpha = 0.754$) and self-referential direction ($\alpha = 0.802$). However, advice strength reliability was poor ($\alpha = 0.371$), indicating that the imperative/recommendation/suggestion distinction is not consistently recoverable from the automated classifications alone. Findings relating to advice strength should therefore be treated as indicative rather than reliable. Factual precision reliability was also below threshold (α = 0.655), though this appears to reflect near-zero variance in the sample rather than systematic disagreement on the vagueness label itself.

Ideally, a full manual re-coding of the corpus would be needed for complete confidence across all dimensions. However, such an undertaking was beyond the scope of this preliminary analysis.

Additional concerns arise for the fact-checking pipeline, as it was not possible to conduct a systematic check for validity in this study. To address this partially, a human auditor independently verified a sample of statements assigned unambiguous labels (Unambiguously true or Unambiguously false) against authoritative sources and validated the result in all cases. However, statements assigned vague or “unable to evaluate” labels were not independently audited, as these resist definitive human adjudication. The overall fact-check results should therefore be treated as indicative rather than definitive, particularly for the vague label categories.

Other Limitations

This is a preliminary study and carries several limitations that constrain the generalisability of its findings. First, the analysis was conducted at a single point in time. LLM responses are not static: the same question may elicit materially different responses across versions or over time (Galadima et al., 2025). The findings reported here reflect the behaviour of these four models as accessed in early 2026, on free-tier accounts with no conversational history.

Second, the question set, while grounded in common search data and financial education frameworks, represents only a small slice of the personal finance questions users might ask.

Questions involving more personal or context-dependent circumstances may elicit different responses.

Finally, this study does not capture the dynamic, multi-turn nature of real user interactions with LLMs. In practice, users often follow up, ask for clarification, or provide context — interactions that may substantially alter the advice received - either for better or for worse.

It bears emphasis that this study is attempting to understand what kind of financial advice the typical unsophisticated LLM user may be receiving, not the potential such tools have for providing personal finance assistance in an optimal use case. It is plausible that modifying the prompts and adding appropriate global instructions for conversations might significantly improve outcomes.

Conclusion

This study provides preliminary evidence concerning the advice received by consumers asking LLMs about personal finance. Several patterns emerge consistently. All four models are predominantly advisory in character, providing dense, action-oriented responses with little effort to gather context about the user's individual circumstances. The volume of advice — typically 10 to 20 items per question — and its directional strength raise questions about whether responses reflect best practices in financial guidance, which emphasize assessing individual needs before recommending a course of action (Theodos et al., 2015). The predominance of leading questions over clarifying ones further suggests that models tend to nudge users toward action rather than seeking to fully understand their situation.

It should also be noted, though, that leading LLMs differ substantially in their responses — not just in length or factual accuracy, but in the rhetorical strategies they use to frame, direct, and contextualise advice. ChatGPT stands out for its heavy use of leading questions and exclusively confidence-oriented self-referential framing, never acknowledging uncertainty or recommending external expertise. Claude, by contrast, asks more clarifying questions, produces fewer and more tentative statements, and is the only model to consistently acknowledge its limitations as a non-specialist source. Grok and Gemini produce longer responses with higher rates of vague or misleading factual claims and greater sensitivity to question wording, suggesting their outputs are less consistent and more susceptible to hallucination in financial contexts.

These findings have implications both for users and researchers. For users there are three general concerns. First, care must be taken to proactively supply the necessary context needed to provide an answer tailored to their individual needs, as the LLM may well make assumptions rather than asking for clarification when providing leading follow-up questions. Second, the user must make sure they are asking precisely the right question, as the LLM will tend to interpret it literally and respond to the exact wording rather than searching for the intention behind it. Third, the user should take care not to be swept along by a large volume of jargon and confident encouragement and make sure they understand everything that's being said.

It is likely possible for a user to work with the model to address these issues through "prompt engineering" - instructing it, for example, to ask for clarification first and only then provide advice. However, this is not *default* LLM use and requires educating users. Better understanding how users can work with LLMs so they provide more effective financial advice is a pressing research topic for future work.

In addition, for consumers the choice of LLM is not neutral from a financial guidance perspective: models differ meaningfully in the confidence they project, the contextual information they seek out, and the accuracy of the factual claims they make.

For researchers, this study demonstrates that systematic content analysis — combining automated classification with targeted inter-coder reliability testing — offers a workable framework for evaluating LLM outputs as communicative acts, not just factual claims.

Several limitations constrain the generalisability of these findings. The study covers five questions asked on a single occasion across free-tier accounts; responses may differ under other conditions, across different phrasings, or as model versions are updated. Advice strength classifications carry insufficient reliability for firm conclusions. Future work should extend this framework to a larger and more diverse question set, incorporate multi-turn interactions that better approximate real user behavior, examine the influence of prompt engineering and examine the downstream effects of these rhetorical patterns on user comprehension, confidence, and financial decision-making.

References

AFCPE (2018). Code of Ethics. https://www.afcpe.org/certification/professional-standards/code-of-ethics/

BBC/EBU (2025, October). AI chatbots misrepresent news content almost half the time. BBC News. https://www.bbc.co.uk/rd/publications/ai-news-accuracy-study

BBC News (2026, February 10). AI chatbots pose ‘dangerous’ risk when giving medical advice, study suggests. https://www.bbc.com/news/articles/ai-chatbots-medical-advice-risk

Chatterji, A., Cunningham, T., Deming, D. J., Hitzig, Z., Ong, C., Shan, C. Y., & Wadman, K. (2025). How people use chatgpt (No. w34255). National Bureau of Economic Research.

Chew, R., Bollenbacher, J., Wenger, M., Speer, J., & Kim, A. (2023). LLM-Assisted Content Analysis: Using Large Language Models to Support Deductive Coding. *arXiv:2306.14924*.

Consumer Financial Protection Bureau (CFPB). (2024). *Your money, your goals: A financial empowerment toolkit*. U.S. Consumer Financial Protection Bureau.

FINRA. (2024). The machines are coming (with personal finance information). Do we trust them? *Consumer Insights* https://www.finrafoundation.org/sites/finrafoundation/files/the-machines-are-coming.pdf

FINRA (2025). The National Financial Capability Survey, 2024. https://www.finrafoundation.org/national-financial-capability-study

Galadima, A., Ngo, V., et al. (2025). Can AI help with your personal finances? *Applied Economics*. https://doi.org/10.1080/00036846.2025.2450384

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1), 77–89.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., … Liu, T. (2025). A survey of hallucination in large language models. *arXiv:2510.06265*.

Inghelbrecht, K., & Tedde, M. (2024). Overconfidence, financial literacy and excessive trading. *Journal of Economic Behavior & Organization*, 219, 152–195. https://doi.org/10.1016/j.jebo.2024.01.010

Kim, K. T., Wilmarth, M. J., & Tian, F. (2020). Financial literacy overconfidence and mortgage delinquency. *Journal of Consumer Affairs*, 54(2), 517–540.

Kim, S. S., Liao, Q. V., Vorvoreanu, M., Ballard, S., & Vaughan, J. W. (2024, June). " I'm Not Sure, But…": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. In *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency* (pp. 822-835).

Krippendorff, K. (2022). *Content analysis: An introduction to its methodology* (4th ed.). SAGE.

Li, Y., Wang, S., Ding, H., & Chen, H. (2024). Large language models in finance: A survey. *arXiv:2311.10723*.

Liberty Bank (2019). The Most Googled Financial Questions by State. https://www.libertybank.com/most-searched-financial-questions/

Lusardi, A., & Mitchell, O. S. (2014). The economic importance of financial literacy: Theory and evidence. *Journal of Economic Literature*, *52*(1), 5–44.

OECD/INFE. (2020). *OECD/INFE 2020 international survey of adult financial literacy*. OECD.

OpenAI. (2026). GPT-5.3 Instant: Smoother, more useful everyday conversations. https://openai.com/index/gpt-5-3-instant/

Tai, R. H., Bentley, L. R., Xia, X., Sitt, J. M., Fankhauser, S. C., Chicas-Mosier, A. M., & Monteith, B. G. (2024). An examination of the use of large language models to aid analysis of textual data. *International Journal of Qualitative Methods*, 23. https://doi.org/10.1177/16094069241231168

Tenet. (2024). *LLM statistics 2026: Adoption, trends, and market insights*. https://www.wearetenet.com/blog/llm-usage-statistics

Theodos, B., Simms, M., Treskon, M., Stacy, C., Brash, R., Emam, D., & Collazos, J. (2015). An evaluation of the impacts and implementation approaches of financial coaching programs. Washington, DC: Urban Institute, 500, 2022-04.

The Currency (2024). Americans’ 10 most-asked money questions. https://www.empower.com/the-currency/money/most-asked-money-questions

Yie, H., Wang, J., et al. (2024). A survey of large language models for financial applications. *arXiv:2406.11903*.

Appendix

Technical Details on Analysis Pipeline

The LLM responses (stored as .docx files) were processed through a multi-step automated pipeline built around the Anthropic API. Rather than splitting responses by character count, documents were chunked by semantic section boundaries — using structural markers including numbered headings, horizontal rules ([HR]), and **bold** section titles. Short adjacent sections under ~200 words were merged to avoid decontextualised fragments. Each chunk was tagged with metadata including question ID, section type (intro, numbered, conclusion, CTA), and section number.

Each chunk was then submitted to Claude (Sonnet) in a separate API call for classification. The coding scheme was developed inductively through iterative testing on sample chunks from the corpus, with the prompt refined in response to observed misclassifications before being applied to the full dataset. This inductive development approach follows recommended practice for LLM-assisted content analysis (Chew et al., 2023).

The final scheme classified every statement into one of six types: factual, advice, normative, question, self-referential, and miscellaneous. Sub-analyses were applied by type:

- **Factual**: precision (precise/vague) and testability
- **Advice**: strength (suggestion/recommendation/imperative), conditionality, and implicit assumptions
- **Questions**: function (clarifying/leading/rhetorical/reflective) and nudge direction
- **Self-referential**: direction (confidence/capability/limitation) and rhetorical function

A key methodological decision was instructing the classification model to treat all analysed responses as coming from an unidentified AI system — this prevented the model from evaluating self-referential claims (e.g., “I’m not a financial advisor”) against its own known capabilities, rather than treating them as analytically opaque assertions by an unknown author.

The prompt also included explicit decomposition rules for compound statements — sentences that simultaneously encode multiple speech acts. The most common patterns addressed were: (1) factual claims backed by high-status authority (“most experts recommend X”), which were split into a factual and an embedded advice statement; (2) named financial frameworks (e.g., the 50/30/20 rule) appearing in advisory sections, split into a definitional fact and an endorsement; and (3) the “I can’t X, but I can Y” construction, always decomposed into a limitation and a capability statement.

Structured JSON output was enforced via the prompt schema, with a six-stage fallback recovery process for malformed responses (markdown fence stripping, prose extraction, bracket-aware truncation recovery). Results were written to CSV.