Judge for Yourself

Confidence Beats Accuracy in AI Financial Advice*

Golnaz Bahrami
Pennsylvania State University

J. Anthony Cookson
Pennsylvania State University

Austin Moss
University of Colorado Boulder

October 9, 2026

Abstract

Millions now ask AI for financial advice. We study the advice they choose on Chatbot Arena, where two randomly assigned, anonymous AI models answer the same question and the user votes for the answer they prefer. Across 3,060 votes on financial questions, users select advice that is confident and actionable. The more accurate advice wins only 52.2% of decisive votes. Even when two paired answers differ greatly in both confidence and accuracy, inaccurate advice that is more confident wins 72.1% of the time. When an answer is verifiable, user choices favor accurate advice, but only modestly. The more accurate advice wins 57.6% of decisive votes when it is possible to verify the advice but only 46.9% of decisive votes if advice depends on the user's circumstances or on the future. When LLM developers train on similar preference data, they can be expected to supply more confident and less accurate advice. In our sample, we find evidence of such catering responses for Meta's Maverick model, which was tuned to perform better on Arena's leaderboard.

*Contact: gxb5298@psu.edu (Bahrami), tony.cookson@psu.edu (Cookson), austin.moss@colorado.edu (Moss).
Preliminary draft. We thank Matt Gustafson, Will Mullins and Marina Niessner for comments. Cookson gratefully acknowledges financial support from the Smeal Endowment at The Pennsylvania State University. All remaining errors are our own.

1 Introduction

Households increasingly consult AI for financial advice (Choukhmane et al. 2026).[1] A household that hires a human adviser can look up the adviser's certification and disciplinary record (Dimmock et al. 2018; Egan et al. 2019; Gerken and Momeni 2026). An AI model has neither credentials nor a track record. As a result, a user of AI must judge its advice from the answer itself. For some questions, an answer provides enough context because a household can verify a quoted contribution limit or follow the calculation for a mortgage payment. For other questions, financial advice depends on a person's circumstances or on the future, and the household might not learn the quality of advice even after acting on it. A household can always see is how confident the answer sounds, and people tend to conflate confidence with expertise (Sniezek and Van Swol 2001; Price and Stone 2004). Yet the most confident advice might not be the most accurate. It is thus an open question whether households reward accurate financial advice and under what conditions.

We study this question on Chatbot Arena, a public platform that asks users to vote on answers to their AI queries, and uses the votes to rank the best LLMs (Chiang et al. 2024). Figure 1 illustrates the user interface. After the user submits a question, the platform picks the two models to answer it, and the LLM remain anonymous until the user votes for one or declares a tie, either both good or both bad. From the standpoint of studying advice, this setting removes the adviser's conflict of interest, because the models, unlike human advisers, earn no commissions and maintain no client relationships. Both answers respond to the same question from the same user at the same moment. A comparison of the two answers holds fixed the question, the user, the language, and the timing, and isolates how users value the content of financial advice.

Chatbot Arena calls each comparison a *battle*. Our sample contains 3,060 battles in which the user asks for financial advice. We downloaded the data from two public releases of the platform's full conversations, spanning two periods, June to August 2024 and April to July 2025.[2] Restricting

[1] Roughly 10% of the world's adult population adopted ChatGPT by mid-2025, with most of its use classified under practical guidance and information seeking (Chatterji et al. 2025). OpenAI and other major platforms have responded to this demand by offering to link users' financial accounts and tailor advice to them.

[2] The data are publicly available to download from Hugging Face.

attention to finance gives a clear benchmark for accuracy, namely the practice standards of the Certified Financial Planner Board and textbook life cycle prescriptions as in Choukhmane et al. (2026). We taksed Claude Sonnet 5 to grade the accuracy of each answer against these benchmarks, and separately, rate how confident and how actionable the answer is. The grader reads one answer at a time and never sees the competing answer, the models' names, or the vote. Treating each battle as a matched pair, we estimate how the vote depends on answers' accuracy versus confidence and actionability.

Figure 2 portrays our main finding. Users reward confidence and actionability while giving accuracy less weight. We call a vote for a single answer *decisive* in contrast to when a user votes for a tie or that both answers are bad. When one answer is more confident, it wins 63.0% of decisive votes. When one answer is more accurate, it wins 52.2%, which is not significantly different from random guessing. In a linear probability model with model and week fixed effects, one-standard-deviation higher confidence is associated with 4.77 percentage points higher choice likelihood. A standard-deviation increase in actionability raises it by 3.85 pp. By contrast, the estimate for accuracy is small and statistically insignificant.

Users reward accuracy when it is easier to verify an answer. The more accurate answer wins 57.6% of decisive votes on questions that are publicly verifiable, such as a contribution limit or the payment on a stated mortgage, and 46.9% on questions that depend on personal circumstances or on the future. From our estimated specification, a one-standard-deviation increase in accuracy is associated with 3.08 pp higher choice likelihood on questions that can be easily verified and -0.82 pp on questions that depend on user considerations or the future. Even when verifiability is high, the preference for accuracy is modest. It is larger for confident answers for which a standard deviation increase in confidence is associated with 4.84 pp higher choice likelihood in the high verifiability subsample. A second test uses Search Arena, a part of the Arena platform in which models answer with live web search and citations (Miroyan et al. 2026). The links and citations make this setting highly verifiable. Consistent with verifiability playing an important role, on Search

Arena, a one-standard-deviation increase in accuracy raises the probability of winning by 7.25 percentage points, while neither confidence nor actionability is related to user choices.

The small accuracy coefficient in the Chatbot Arena sample reflect noise in measuring accuracy. The verifiability results argue against a simple measurement error critique, since the same scores predict the vote when users can check the answer. The scores also predict how users judge a pair as a whole. A one-standard-deviation increase in the average accuracy of the two answers significantly reduces the probability that the user rates both as bad by 4.61 percentage points. A related concern is that the confidence coefficient picks up part of the reward for accuracy. Sorting battles by the accuracy gap leaves the confidence premium about the same whether the more confident answer is the less accurate or the more accurate of the two. When the two attributes point in opposite directions, users tend to favor the more confident answer. An answer that is at least two points more confident and at least two points less accurate than its pair wins 72.1% of decisive votes.

Beyond the platform, these findings matter because revealed preferences feed into model training and LLM leaderboard rankings (Ouyang et al. 2022; Singh et al. 2025). Trained models effectively cater to user preferences. We investigate this supply-side response using an Arena leaderboard scandal. In April 2025, Meta entered a non-public experimental version of its Llama 4 Maverick model to the Arena, which was optimized to cater to human preferences.[3] This experimental version of Maverick won 62.3% of its decisive votes on financial questions in comparison to 34.4% for the public release of the same model. In battles in which the experimental release and the public release faced off, the experimental version won 63 of the 83 decisive votes. Relative to the public release, the experimental model's financial answers were 0.71 points more confident, 1.14 points more actionable, and 0.56 points *less* accurate, and for low-verifiability questions, the confidence gap was especially large. The comparison shows how a model can successfully cater to users on financial questions – by increasing confidence and actionability while neglecting accuracy.

[3]The platform's statement is quoted in Tobias Mann, "Meta accused of Llama 4 bait-and-switch to juice LMArena rank," *The Register*, 8 April 2025. Llama 4 Maverick's experimental version only a small subset of our financial advice sample, contributing 77 out of the 2,212 decisive votes in our sample. Dropping this model's battles, the coefficients in column (1) of Table 2 are 5.14 percentage points for confidence, 3.80 for actionability, and 0.99 for accuracy, very similar to our main results.

Our primary contribution is to the literature on financial capabilities of households (Campbell 2006; Lusardi and Mitchell 2014; Keys et al. 2016). We specifically relate to work that asks why households fail to act on sound financial advice (Bhattacharya et al. 2012; Reuter and Schoar 2024). Existing explanations focus on trust (Gennaioli et al. 2015; Gurun et al. 2018), conflicts of interest (Inderst and Ottaviani 2012; Mullainathan et al. 2012; Anagol et al. 2017; Foerster et al. 2017; Egan et al. 2019; Barbu and Wang 2025; Sun and Schoar 2024), and the supply side of financial advice (Linnainmaa et al. 2021). Our contribution is to investigate households preferences for which advice to take. The Chatbot Arena setting trust and conflicts of interest as candidate explanations because the adviser is anonymous and earns no commission. We still find that households prefer confidence while neglecting accuracy. A key implication is that poor advice can persist through the client's choice rule, much as in markets where buyers weigh the attributes they can see and neglect those they cannot assess (Gabaix and Laibson 2006; Bordalo et al. 2013; Cookson et al. 2026a).

We also contribute to research on AI's use in household finance. Recent work has shown that recommendations by robo-advisers and LLMs can benefit households if they follow their prescriptions (D'Acunto et al. 2019; D'Acunto and Rossi 2023; Reher and Sokolinski 2024; Rossi and Utkus 2024; Fieberg et al. 2025; Choukhmane et al. 2026). However, the potential gains depend both on whether households trust the technology enough to adopt it, and whether households can distinguish good machine advice from bad (Dietvorst et al. 2015; Logg et al. 2019; Chak et al. 2022; Shekar et al. 2025; Okat et al. 2025). Choukhmane et al. (2026) show that the way a household phrases its question shapes the objective quality of AI financial advice; however, they do not study adoption of this advice. We complement this work by investigating a complementary component of the quality of AI financial advice. In our setting, prompt phrasing is held constant but two choices are presented to users, often with meaningful differences in accuracy. By finding that accuracy is not rewarded in these choices, our paper casts doubt that households can mitigate inherent quality differences in AI financial advice by filtering out the good from the bad advice.[4]

[4]Related to our work, Moss et al. (2026) shows that investors steer AI recommendations toward riskier and less diversified portfolios. We complement these insights by studying a choice between competing AI advice where it can be objectively scored on accuracy, rather than catered to risk preferences.

One implication of households favoring confidence over accuracy in financial advice is that this can distort the financial advice that is supplied. Existing professional levers like best-interest standards and fiduciary duty can influence adviser behavior (Egan et al. 2022; Bhattacharya et al. 2025), but if clients choose confident and inaccurate advice, advisers who compete for their business have an incentive to supply poor advice. In lab experiments, Radzevick and Moore (2011) showed that competition could force advisers to trade accuracy for confidence. In the age of AI supplied financial advice, these revealed preferences have an even more direct feedback loop: developers train using human preference ratings (Ouyang et al. 2022), which tend to favor assertive and agreeable answers on general tasks (Hosking et al. 2024; Sharma et al. 2024). Reinforcing this feedback loop idea, in sample, we document how one supplier responded to these preferences. Meta's experimental release of the Maverick model gained votes by systematically tilting its responses to be more confident and actionable while becoming less accurate.

Finally, we contribute to the literature on credence goods (Darby and Karni 1973) in which buyers know little about the quality of the good, and thus, are susceptible to exploitation (Dulleck and Kerschbamer 2006; Dulleck et al. 2011). Financial advice is often given as a leading example of a credence good (Kerschbamer and Sutter 2017). Running counter to this narrative, recent work on the democratization of finance suggests that investors can judge ideas for themselves, since crowd-sourced research and social media signals are informative about returns, earnings, and corporate decisions (Chen et al. 2014; Jame et al. 2016; Farrell et al. 2022; Bradley et al. 2024; Cookson et al. 2024, 2026b). Our findings are consistent with users being unable to assess the accuracy of financial advice, and thus, support the interpretation that financial advice is a credence good. However, when advice is verifiable, either through its content or through source documentation, we find evidence that users are able to recognize and reward accuracy. Whether advice behaves as a credence good therefore depends on whether the structure of advice allows the client to verify the answer.[5]

[5]In a general credence goods setting, Sah et al. (2013) provide evidence that confident inaccurate advice is punished if it is verifiable, but rewarded when it is not verifiable. Our work takes this evidence into the field and into the setting of financial advice.

2 Data and Measurement

This section describes the platform and the votes it records, the sample construction, a battle that illustrates what we measure, how we score each answer, the verifiability label, and the within-battle differences that serve as our regressors. Appendix A lists the variable definitions.

2.1 The Arena Setting

Our data come from Chatbot Arena, a public website on which a user types a question, receives answers from two LLMs, and votes for one answer, declares a tie, or records that both answers are bad (Chiang et al. 2024). Figure 1 follows one battle through these steps on the current version of the site. The user types a question in Panel A, and the platform assigns two anonymous models to answer it. Panel B shows the two answers side by side under the labels Assistant A and Assistant B, with the voting buttons beneath them. The platform reveals which model wrote each answer only after the user votes, as in Panel C.

A battle consists of one question, its two answers, and the vote. We call the two answers side A and side B according to their positions on the screen. A *decisive vote* is a vote for side A or for side B, as opposed to a tie or a both-bad vote. The position of an answer carries no information about its quality. In our sample, side A wins 48.4% of the decisive votes, close to one half.

Two features of the platform make the vote a comparison between two answers alone. First, the user neither chooses the two models nor knows which models wrote the answers at the time of the vote. In our anonymous battles, the platform picks two models and hides their names until the user votes, as Panels B and C of Figure 1 show. The user therefore cannot favor an answer for the brand or reputation of the model behind it.[6] A separate side-by-side mode lets the user pick two models and see their names, but those votes do not count toward the platform's public leaderboard and are not available in the public releases we accessed. In the anonymous mode, the platform also picks

[6]The platform's FAQ states that votes in the side-by-side mode do not contribute to the public leaderboard rankings (https://arena.ai/faq, accessed October 2026).

the pair independently of the question. A user with a hard question is as likely as a user with an easy question to draw a given pair of models at a given time.[7]

Second, both answers respond to the same question, from the same user, at the same moment. Everything about the asker and the question is therefore common to both answers, including the question's difficulty, what the asker already knows, the money at stake, the language, and the time of day. The within-battle comparison holds these factors fixed, and it lets us study how users weigh the content of the answers themselves.

Chatbot Arena users have no financial stake in the vote and may be more engaged with technology than the average household. Even so, their votes determine the public leaderboard rankings, and developers who compete for those rankings have an incentive to cater to these users. Some votes are likely inattentive or adversarial, and such votes can shift the rankings (Zhao et al. 2025). Inattentive votes, cast without regard to the answers, would shrink every coefficient in our vote regressions toward zero in the same proportion. Because proportional shrinkage leaves the ratio of any two coefficients unchanged, it cannot by itself make the confidence coefficient larger than the accuracy coefficient.

2.2 Sample Construction

The core sample consists of the 3,060 battles in which the user asks a finance question for advice or for information. We draw them from the 241,768 voted battles with full conversations in two public releases of Chatbot Arena data on Hugging Face.[8] The 2024 release contains 106,134 battles from 20 June to 19 August 2024, and the 2025 release contains 135,634 battles from 17 April to 24 July 2025. A keyword filter followed by an LLM classifier selects the finance battles. Because neither step reads the answers or the vote, selection does not use the outcome of a battle.

The keyword filter marks a battle as a finance candidate when the user's messages contain any of 103 English finance terms or any of 44 terms in German, French, Spanish, Russian, Chinese,

[7]Every battle in the 2024 release is flagged as anonymous, and the 2025 release documentation states that new models are sampled after each vote, which is how anonymous battle mode works. Sampling rates vary across models and over time (Chiang et al. 2024; Singh et al. 2025).

[8]The 2024 release is the Hugging Face dataset `lmarena-ai/arena-human-preference-100k`, and the 2025 release is `lmarena-ai/arena-human-preference-140k`.

Japanese, and Korean. The English terms include "mortgage," "retirement," "401(k)," and "inflation." We intentionally keep general words such as "bank," "income," and "fees" on the list so that the filter stays broad, and we rely on the classifier to remove battles that match a keyword by mistake. A finance question that uses none of the terms is not in the sample. In particular, a question in a language other than these eight enters the sample only when it contains one of the listed terms. The platform's language tag places 229 of the 3,060 core battles in such languages. The keyword filter selects 13,784 candidate battles, 6,716 from the 2024 release and 7,068 from the 2025 release.

The classifier, Claude Haiku 4.5, reads the first 600 characters of the same messages and assigns one of five labels, two of which define the core sample. An *advice-seeking* question describes or clearly implies the user's own financial situation or decision, such as whether to invest, save, borrow, or buy rather than rent. The question in Figure 3 is advice-seeking because it asks what house the user can afford given the their income and cash. An *information-seeking* question asks about finance without stating a personal decision, as in "what is a carry trade in financial markets." A question about whether a stock is a good buy, asked with no personal context, is also information-seeking.

We drop the battles with the other three labels. A task asks the model to produce work on a finance topic, such as an essay, code for a trading program, or the answer to a homework problem, and the vote judges that work, not the advice the user would act on. A test shows explicit signs that the user is probing the model, such as comparing AI models, asking a trick question, or trying to make the model break its rules, and the vote records which model passed. A battle is not about finance when its prompt matches a keyword but does not concern personal finance, investing, or financial markets, such as the bank of a river or bonds in chemistry. Of the 13,784 candidates, the classifier labels 5,488 as tasks, 417 as tests, and 4,819 as not about finance. The remaining 1,323 advice-seeking and 1,737 information-seeking battles form the core sample.

The 2024 release contributes 907 core battles, of which 294 seek advice and 613 seek information, and the 2025 release contributes 2,153, of which 1,029 seek advice and 1,124 seek information. The core battles include 106 distinct models, and 74.6% of the battles are in English according to the platform's language tag. In 18.5% of these battles, the user writes more than one message, and

we score only the first answer on each side. Of the 3,060 votes, 2,212 are decisive, with 1,071 for side A and 1,141 for side B. The rest are 529 ties and 319 both-bad votes.

We also use Search Arena, a version of the platform in which both models answer using live web search and return citations that let users check an answer against its sources (Miroyan et al. 2026). Of the 24,069 conversations in its public release, 12,652 end in a vote, and the same keyword filter selects 800 of the voted conversations. Claude Sonnet labels these 800 conversations with the same instructions the Chatbot Arena classifier follows. Of the 800, 77 are advice-seeking, 224 are information-seeking battles, 118 are tasks, 8 are tests, and 373 are not about finance. Because the Search Arena sample is small, we keep all 427 battles with one of the four finance labels, including the tasks and tests. These battles run from 18 March to 8 May 2025, and 297 of them end in a decisive vote.

2.3 An Example Battle

Figure 3 reproduces a battle from the 2024 release in which the user voted for an answer that contradicts its own arithmetic. The user asks for the highest house price they could afford on an income of $100,000 a year with $150,000 in cash, assuming a 6% interest rate. Both answers cite two public lending guidelines. The first caps housing costs at 28% of gross monthly income, and the second, a debt-to-income (DTI) ratio limit, caps total debt payments at 36% of the same income. The user chose side A, which commits to a number. It proposes a $650,000 house with a $500,000 loan, sets total monthly housing costs at $4,081, and concludes that this cost "is higher than the 28% guideline, but within the 36% DTI ratio." Its own figures say otherwise because a cost of $4,081 against a gross monthly income of $8,333 is 49% of income, above both caps. Side B opens with "I can't give you an exact maximum price for a house," lists what a lender would look at, states the same two guidelines, and tells the user how to run the calculation.

The battle shows the two dimensions we measure. The first is the answer's confidence and actionability, which can be high or low, whether or not the answer is accurate. In this battle, the more confident and more actionable answer is the less accurate one. The second is the verifiability of the question. A reader can check the house question with arithmetic and the public lending

guidelines. By contrast, many questions in our sample depend on the asker's circumstances or on the future, and public references alone cannot answer them.

2.4 Measuring What Each Adviser Says

An LLM grader, Claude Sonnet 5, scores each answer separately. It sees the user's first prompt and one answer, but never the competing answer, either model's identity, or the vote.[9] As a result, a within-battle difference in a score measures how the two answers differ on that attribute and reflects no judgment by the grader about which answer should win.

The grader scores three attributes. *Confidence* measures how assertive and definitive the answer is, regardless of whether it is right. The scale runs from 1, for an answer that refuses to engage or is wrapped in uncertainty, to 5, for definitive recommendations or predictions with essentially no hedging. *Actionability* measures how concrete and implementable the guidance is. It runs from 1, for an answer with nothing the user could act on, to 5, for a complete and specific implementation with ordered steps, named products or account types, and dollar or percentage figures.

In the battle of Figure 3, side A scores 4 on confidence because it commits to a number and qualifies it only in a closing caveat, and 5 on actionability because it gives a complete breakdown the user could act on. Side B scores 2 on confidence because it declines to name a price and hedges throughout, and 3 on actionability because it gives the guidelines and the procedure but does not run the calculation.

The third attribute, *Accuracy*, measures whether the answer is correct and suits the person who asked. The user sees confidence and actionability on the page but often cannot judge accuracy from reading alone. In a second pass over every answer, the grader uses a written rubric to score two subscores, factual correctness and suitability, and combines them into the accuracy composite. *Factual correctness* grades the answer's checkable claims about definitions, mechanics, tax rules, product features, named entities, and arithmetic. The scale runs from 1, when the answer's core factual content is wrong, to 5, when every claim is correct.

[9] Each item consists of the first 4,000 characters of the prompt and the first 6,000 characters of the answer, and the two answers of a battle are never placed in the same batch. Of all answers, 11.0% exceed the 6,000-character limit. The instructions given to the grader are reproduced in the replication package.

Three rules shape the factual correctness subscore. First, fabrication caps the subscore at 2. Fabrication means naming a product or entity that does not exist, presenting a defunct one as current, quoting a precise figure as data with no source, or stating numbers that contradict one another. A claim that is stale, rounded, or true in only some cases is an ordinary error and by itself lowers the subscore no further than 3. Either kind of error counts against the answer only when the grader is confident that the claim is false. Second, an answer with no checkable claim, such as a bare refusal, scores a neutral 3. Third, a prediction or an opinion lowers factual correctness only when the answer presents it as a fact.

The second subscore, *suitability*, grades whether the advice fits the circumstances and goals the asker describes. Suppose a user with a large, high-interest credit card balance asks which index fund to buy. An answer that describes a low-cost index fund correctly but never mentions the debt can score 5 on factual correctness, yet only 3 on suitability because the user would do better to pay off the card first. When the grader judges that the question involves no personal decision, suitability grades whether the explanation is complete and calibrated. Such an explanation answers what was asked, includes the caveats a correct understanding requires, and claims no more than is known.

Grading fit requires a benchmark other than the grader's own view of what the asker should do. We use the standards certified human planners must follow as this benchmark. The Certified Financial Planner Board of Standards (CFP Board) awards the CFP certification in the United States. Its practice standards for the financial planning process, in force since October 2019, require a planner to understand the client's personal and financial circumstances, identify the client's goals, and analyze the current course of action and its alternatives before presenting a recommendation. Because these standards guide how a planner reaches a recommendation, the grader can apply them even when the right choice depends on facts the asker has not given. The grader cannot know whether a particular asker should pay off a mortgage or invest, but it can tell whether the answer raises the facts that decide the matter, such as the asker's time horizon, emergency fund, other debts, and tax situation. For personal decisions, the rubric also asks whether the recommendation aligns with the life-cycle prescriptions Choukhmane et al. (2026) use as a benchmark, such as broad participation in low-cost, diversified equity funds, an equity share that declines with age, and a liquid savings buffer before speculative positions.

The composite starts at the lower of the two subscores. The grader can move it up or down by one point, but never above 2 when the answer fabricates a fact. The rubric tells the grader to score accuracy independently of confidence, length, and formatting. A hedged answer can score 5, and a confident, polished answer can score 1. However, two suitability rules link accuracy to confidence, one at each end of the confidence scale. A confident answer to a question that cannot be answered as posed scores at most 3 on the composite, a cap that penalizes overconfidence. A bare refusal, the least confident kind of answer, faces the same cap because it gives the user nothing to use. An answer that declines to give specifics but provides a sound way to decide, such as side B in Figure 3, is not a bare refusal, and the grader scores it on what it provides.

Side A in Figure 3 scores 1 on factual correctness because its core factual claim is wrong. Its payment arithmetic is right, but its $650,000 answer rests on calling a housing cost of 49% of income within the 36% limit it has just quoted. Side A scores 2 on suitability, the rubric's score for advice that fails on the main margin of the decision, which here is affordability. The composite of 2 is the lower subscore plus the one-point adjustment. Side B scores 5 on both subscores and on the composite. By listing what a lender would look at, it raises the facts that decide the matter.

Each of the five scores takes every value from 1 to 5 in the data. The mean is 3.70 for confidence, 3.02 for actionability, 3.63 for accuracy, 3.69 for factual correctness, and 3.99 for suitability. To check the grader's consistency, we have it score 250 randomly drawn answers a second time in both passes.[10] We measure agreement between the two scorings with the quadratic weighted kappa, which weights each disagreement by the square of its size and equals 1 under perfect agreement and 0 under chance agreement. The kappa is 0.81 for confidence, 0.87 for actionability, 0.74 for accuracy, and 0.75 for factual correctness. It is lowest, 0.69, for suitability, which asks the grader to judge fit rather than check claims. Appendix Table A2 reports the means and lists each kappa as the two-pass reliability of the score. Measurement error in LLM ratings tends to be milder for shorter texts (Cookson et al. 2026c), and the answers we score are short, with a median of 399 words in English tagged battles.

[10]The second scoring of confidence and actionability returns scores for 249 of the 250 answers, and the kappas for those two attributes use these 249.

2.5 Measuring Verifiability

The *verifiability label* measures how far a diligent reader could check an answer to the question against public references without knowing more about the asker. Claude Sonnet assigns the label in a separate pass with its own instructions. It reads only the first 600 characters of the user's first message.[11] The labeling pass never sees the answers, the grader's scores, or the vote. We keep the answers out on purpose. A label that saw them could call a question checkable because one of its answers is plainly wrong.

The label takes three values. We call a question *high* when its answer is a matter of public record or mechanics, such as a definition, a tax or contribution rule, the result of a formula, or a historical fact. Examples in the instructions include how a Roth IRA differs from a traditional IRA, the 2024 401(k) contribution limit, and the monthly payment on a $300,000 mortgage at 6%. A *medium* question has a checkable factual core but also calls for judgment. Whether ETFs or mutual funds are better is one, and facts constrain the answer without settling it. A question about whether now is a good time to buy bonds is medium too, because the facts can be checked today but the asker usually wants an interpretation. A question is *low* when its answer depends on the asker's circumstances or preferences, or on the future. Personal decisions such as whether to pay off a mortgage or invest fall here, along with predictions such as whether Tesla stock will go up, opinions, and questions that cannot be answered as posed.

Of the 3,060 core battles, 1,376 are labeled high, 617 medium, and 1,067 low. Users cast a decisive vote in 952, 476, and 784 of them, respectively (Table 3, Panel A). In the same pass, Claude Sonnet identifies the 386 battles whose question asks for a forecast of prices, returns, rates, or other economic outcomes. No reference can check such a forecast when the user votes. None of the 386 is labeled high.

Verifiability is correlated with question type. Advice-seeking questions make up 21.6% of high-verifiability battles, 41.3% of medium-verifiability battles, and 72.3% of low-verifiability battles (Table 3, Panel A, third column). A question about the asker's own finances usually depends on facts that only the asker knows. The correlation is imperfect. The house question of Figure 3

[11] When a message contains several questions, the label applies to the dominant one. Messages in other languages follow the same rules. The replication package reproduces these instructions.

seeks advice, yet Claude Sonnet labels it high. Its asker states the income, the cash, and the interest rate, and the answer follows from arithmetic on those numbers and from public lending guidelines.

2.6 Differences and Pair Averages

Each battle gives us two answers to the same question. Our regressors compare them. For a variable x measured on an answer, write x_A and x_B for its values on the side A and side B answers. The *within-battle difference* is

$$\Delta x \;=\; x_A - x_B,$$

which is positive when side A scores higher. In the battle of Figure 3, side A is more confident, more actionable, and less accurate than side B. In raw points, the differences are +2, +2, and −3.

We divide each difference by s_x, the standard deviation of Δx over all 3,060 core battles, and call the result the standardized difference. We standardize because our main test compares the weight users put on confidence with the weight they put on accuracy. A one-point gap does not mean the same thing on the two scales, and accuracy gaps are more dispersed than confidence gaps. Standardizing changes only the units. To get the coefficient per one-point gap, divide the coefficient per standard deviation by s_x. The divisors are 1.030 for confidence, 1.131 for actionability, and 1.200 for accuracy. A one-standard-deviation gap is thus close to a one-point gap on the 1 to 5 scale. Appendix Table A3 also lists the divisors for the two accuracy subscores and for the log of answer length.

We compute the divisors once, on the core sample, and use them for every Chatbot Arena subsample. A coefficient from a subsample then refers to the same gap in raw points as a coefficient from the core sample. Tables report standardized differences unless a heading says raw units.

The *pair average* of each attribute is the average of its scores on the two answers in the same battle,

$$\bar{x} \;=\; \frac{x_A + x_B}{2}.$$

For the battle of Figure 3, accuracy scores of 2 and 5 give a raw pair average of 3.5. Before a pair average enters a regression, we center it and divide it by its standard deviation over the 3,060 core battles.

The difference records which answer scores higher, and the pair average records how good the pair is. Between them they carry the same information as the two scores, since $x_A = \bar{x} + \Delta x/2$ and $x_B = \bar{x} - \Delta x/2$. A regression that includes both uses everything the grader reports.

3 Main Results

In this section, we begin by describing our main results on the Chatbot Arena sample, starting with the empirical specification. After describing these main results, we examine the role of verifiability and its interaction with accuracy and confidence. Complementing the textual measure of verifiability, we analyze preferences on Search Arena where answers are naturally more verifiable due to inline citations.

3.1 Empirical Specification

Our main specification considers only decisive battles in which either side A or side B wins. Votes for a tie or a both-bad vote are not included in the main sample, leaving only the 2,212 decisive votes. Section 4.1 considers alternative dependent variables in which user votes are not decisive.

For battle b in calendar week $w(b)$, we estimate a linear probability model[12] of the vote on the three standardized differences of Section 2.6

$$\begin{aligned} \text{Win}_b = \beta_1 \frac{\Delta\text{Confidence}_b}{s_{\text{conf}}} + \beta_2 \frac{\Delta\text{Actionability}_b}{s_{\text{act}}} + \beta_3 \frac{\Delta\text{Accuracy}_b}{s_{\text{acc}}} \\ + \sum_m \gamma_m D_{mb} + \delta_{w(b)} + \varepsilon_b. \end{aligned} \tag{1}$$

where the dependent variable Win_b equals 100 if answer A wins and 0 if answer B wins. We employ this coding so that the β coefficients reflects the percentage point change in the likelihood that

[12] Only 3.0% of the fitted win probabilities fall below 0 or above 100, and the average marginal effect of a logit lie within 0.33 percentage points of the linear coefficients (Appendix Table C6).

side A wins due to a one unit movement of an attribute. We scale each attribute (confidence gap, actionability gap, and accuracy gap) by its in-sample standard deviation, so the coefficient reflects the shift in the choice likelihod for a one standard deviation move.

We are interested in β_1, β_2 and β_3 because these describe whether a difference in confidence, a difference in actionability and a difference in accuracy predicts the user's choice of financial advice. A key test contrasts the coefficients on accuracy with confidence, implemented as a Wald test that users put equal weight on confidence and accuracy, $\beta_1 = \beta_3$. Table 2 reports its p-value in brackets.

The model fixed effects are signed indicators. For each of the 106 models, D_{mb} equals +1 when model m writes the side A answer in battle b, -1 when it writes the side B answer, and 0 otherwise. Its coefficient γ_m measures the general appeal of model m to voters, and the predicted vote depends on the difference between the two models, as in the Bradley–Terry form of paired comparisons. Fixed effects $\delta_{w(b)}$ for the 25 calendar weeks absorb side A's win rate in each week.

Because the platform picks the two models without considering the question and the LLM grader scores each answer without seeing the vote, we interpret the coefficients as reflecting the weight a user puts on each attribute when choosing between two answers. The error ε_b captures features voters value, but the LLM grader does not score, such as length, formatting, and tone, and these features can move with the scored gaps. Model indicators absorb any model-specific features, and each β comes from variation in an attribute gap that the identities of the two models do not explain. Standard errors are heteroskedasticity-robust with the HC1 small-sample correction unless a table note says otherwise.

3.2 Descriptive Evidence

Table 1 splits the decisive votes by which answer scores higher on each attribute, and Panel B reports the share won by the higher-scoring answer among votes in which the two answers differ. More confident answers win 63.0% of these votes, and more actionable answers win 65.0%. For accuracy, the share is 52.2%, not significantly more than half. Answers with the higher suitability subscore win 55.6%, and longer answers wins 63.9%.

Side A's win rate in Figure 2 increases as the confidence gap (Side A minus Side B) increases. It is 15.0% when side A is three or more points less confident than side B, and it increases 75.6% when it is three or more points more confident. The accuracy profile stays between 47.7% and 52.4% for gaps from −2 to +2 and departs from an even split only when side A is three or more points less accurate, where side A wins 31.9%.

The flat accuracy profile is not due to there being not much variation in accuracy between the two answers. In Figure 4, all three gaps are symmetric around zero, and the grader gives the two answers the same score in roughly half of the battles for accuracy, confidence and actionability. In fact, accuracy gaps vary the most, with a standard deviation of 1.20 versus 1.03 for confidence.

Most answer pairs are confident In the bottom row of the figure, pair averages of confidence are typically 4 or higher. Pair averages of actionability spread almost evenly across the scale, and those of accuracy peak at 4 with a pronounced left tail.

3.3 Regression Evidence

In this section, we formalize the descriptive evidence from the previous section by reporting estimates from the linear probability model in equation (1). Estimates from this main specification are reported in Table 2.In column (1), a one-standard-deviation increase in the confidence gap is associated with a 4.77 percentage points higher probability that side A wins and the corresponding estimate for actionability is 3.85, both significant at the 1% level. The accuracy is 1.04 and statistically insignificant, and a Wald test rejects equality of the confidence and accuracy coefficients ($p = 0.009$).[13]

We find no evidence that accuracy predicts choice likelihoods more strongly for advice-seeking questions, where the user arguably faces a personal decision. In column (3) of Table 2, confidence and actionability remain significant at the 5% level, and the accuracy coefficient is −0.73 and statistically insignificant.

[13]Without the model indicators, every coefficient is large. In columns (1) and (2) of Appendix Table C3, the coefficients are about 7.2 percentage points for confidence, 8.2 for actionability, and 4.0 for accuracy, all significant at the 1% level.

The confidence coefficient is significantly larger than the coefficient on two distinct components of accuracy. Columns (4) and (5) replace the accuracy composite with factual correctness and with suitability in turn. Factual correctness has a small and statistically insignificant coefficient of 1.00 percentage point, and suitability has a coefficient of 2.04, significant at the 10% level.

Estimated separately by release in columns (6) and (7), the confidence coefficient is positive and significant in both, 8.45 percentage points in 2024 and a smaller 2.99 in 2025. Actionability has a larger coefficient in 2025 than in 2024.

Column (2) of Table 2 clusters standard errors on the 1,162 unordered model pairs. A user who votes several times could also bring the same tastes to each vote. The last two rows of Table 6 cluster the 605 decisive votes of the 2024 release by their 520 voters. Without a voter identifier in the 2025 release, we cluster its 1,607 decisive votes by their 1,529 evaluation sessions. Confidence remains significant at the 5% level or better in both releases.

3.4 Verifiability and the Role of Accuracy

If users do not select more accurate answers because they cannot tell which answer is accurate, the coefficient on accuracy should strengthen as the answer becomes more verifiable. We test this prediction in Table 3 and Figure 5, using a verifiability label classified from the question text.

Panel A of Table 3 reports raw win rates, split by our verifiability measure. On high verifiability questions in which accuracy is different,the more accurate answer wins 57.6% of battles. The rate falls to 50.6% on questions with medium verifiability and to 46.9% for questions with low verifiability. For both low and medium verifiability, these win rates are statistically indistinguishable from 50%. As verifiability decreases, more confident answers are more likely to be chosen from 60.7% with high verifiability to 66.1% with low verifiability. Figure 5 shows that the accuracy profile slopes upward only when verifiability is high, while the confidence profile slopes upward irrespective of verifiability.

Panel B estimates equation (1) separately at each level of verifiability. This lets the model and week effects differ across verifiability levels along with the attribute coefficients. With this more

flexible specification, the estimates have the same ordering, with a positive accuracy coefficient on high-verifiability questions and a negative one on low-verifiability questions.

3.5 Search-Grounded Answers

Search Arena offers a setting in which users can verify their answers easily because both models answer with live web search, and the answers include citations (Miroyan et al. 2026). Thus, we use Search Arena battles as a high verifiability subsample. From the same source as our main Chatbot Arena sample, we have access to 427 Search Arena battles where the question is about financial advice.

In Table 4, we present linear probability model estimates using the 297 Search Arena battles in which a decisive vote was cast by the user. The results contrast strongly with the main sample of Chatbot Arena battles in that users value answers that are more accurate, and do not seem to favor more confident or more actionalbe answers. We find that a standard deviation increase in the accuracy gap (Answer A minus B) is associated with 7.246 percentage points higher likelihood that the user selects answer A. Unlike the accuracy, we find a statistically insignificant relation between the confidence gap and the choice likelihood, and similarly insignificant for actionability. As illustrated in Table C6, the results are similar when employing a logit specification.[14]

In addition to these main measures, we also employ a measure for the number of citations, which count the number of inline citation markets in an answer's text plus the number of sources cited by the answer. Throughout Table 4, we find support that answers with more citations are favored by users, with the coefficient particularly strong in column (3)'s specification with week fixed effects. A standard deviation increase in the citation gap is associated with 11.251 pp greater choice likelihood, significant at the 1% level.

[14]Column (5) of Table 4 keeps the 209 decisive votes on advice-seeking and information-seeking questions, which drops financial tasks from the sample. In that smaller sample, we find qualitatively similar results, though the actionability gap becomes statistically significant at the 10% level.

4 Mechanisms, Validation and Extensions

In this section, we run several analyses that validate our measurement and probe mechanisms. We next consider several robustness exercises, and we conclude the section by examining an episode that illustrates how suppliers cater to user preferences, exemplified by Maverick model developed by Meta.

4.1 Measurement Validation, Relating to Pair Averages

This section estimates specifications that predicts non-decisive votes, which helps to validate our measurement and specification choices.

First, in Panel A of Table 5, we relate confidence gaps, actionability gaps and accuracy gaps to three different outcomes for the full sample of 3,060 financial battles, including non-decisive votes. The dependent variable equals 100 when the vote is a side A win in column (1), a tie ("both good") in column (2), and a both-bad vote in column (3). Column (1) is consistent with our main findings, with significant positive loadings on confidence gaps and actionability gaps, but a small coefficient on the accuracy gap. Because this specification treats non-decisive votes as 0, this column illustrates that our main findings are not driven by how we define a preference for side A.

Turning to non-decisive outcomes (ties and both-bad votes), the estimated coefficients are smaller and insignificant, except for the confidence gap in the both-bad regression. In this case, a standard deviation larger confidence gap predicts a 1.91 pp lower likelihood of a both-bad vote. These generally smaller and insignificant loadings for non-decisive votes make sense because non-decisive votes explicitly do not state a preference between the two models.

In Panel B, we also include pair averages (e.g., the average level of confidence across the two competing answers) in the specification. A higher pair average should not favor selecting side A over side B, but if the underlying constructs are well measured, pair averages should tightly predict the likelihood the user places a non-decisive vote. Moreover, we can expect an increase in the pair average to predict an increase in likelihood of ties ("both good") and decrease the likelihood

of both-bad votes. For both accuracy and confidence, this is precisely the pattern we find. For a standard deviation increase in confidence, we predict a 2.58 pp increase in the likelihood of a tie and a 2.38 pp decrease in the likelihood of a both-bad vote. Similarly, a standard deviation increase in accuracy predicts a 2.8 pp increase in a tie vote and a 4.61 pp decrease in a both-bad vote. Unlike accuracy and confidence, which clearly reflect the quality of answers, we estimate small and insignificant coefficients on actionability.

4.2 Does Confidence Proxy for Accuracy?

In this section, we consider the interaction between accuracy and confidence, as well as the interaction between accuracy and actionability. In a descriptive test in Figure 6, we repeat the exercise of plotting win rates by confidence gap, split by whether accuracy is the 2 or more points lower for side A (much less accurate), 2 or more points higher for side A (much more accurate) or of similar accuracy (-1, 0 or +1 accuracy gap). Irrespective of the level of accuracy, we find that greater confidence increases the choice likelihood about the same. In the right panel, we see a similar gradient for actionability gaps at different levels of confidence.

Quantitatively, even for decisive battles where side A is at least two points less accurate, its win rate still rises from 6.7% when side A is 2+ points less confident to 65.7% when side A is 2+ points more confident, a difference of 59.0 pp, compared with 66.5 pp among the battles in which side A is at least two points more accurate. Actionability shows a similar pattern.

Appendix Table C2 reports the same comparison in regression form. The outer bins are too small to estimate the 106 model indicators, so Panel A of that table replaces them with a single control and keeps the week fixed effects. For each release, we estimate each model's Bradley–Terry strength from the decisive votes in non-finance battles. We call the standardized version of this variable *out-of-domain strength* because it measures a model's general appeal on non-finance battles. Column (7) of Appendix Table C3 applies it to all 2,212 decisive votes. Controlling for out-of-domain strength does not change the coefficient estimates much. However, out-of-domain strength differences between the

models is a strong predictor of the choice likelihood. A standard deviation increase in out-of-domain strength gap (side A minus side B) is associated with 15.36 pp higher likelihood of choosing side A.

4.3 Confident Inaccurate Advice

Figure 7 displays the joint distribution of confidence gaps, accuracy gaps and actionability gaps to illustrate how often they point in opposite directions. We label an answer a confident and inaccurate when it is 2+ points more confident than its pair and 2+ points less accurate. Using this definition, there are 83 battles with an answer that is confident and inaccurate, 2.7% of the core sample. On this subset, the confident inaccurate answer wins 72.1% of decisive battles. The actionable inaccurate answer is about as common and wins 74.2% of its decisive votes. This rate is not entirely driven by battles in which the other side refuses to answer.[15]

Confident inaccurate battles outnumber non-confident accurate by a ratio of 2.8 to one. In these rare cases where confidence and accuracy are much different and align, the vote is close to unanimous. The answer that is both more confident and more accurate wins 90.9% of its decisive battles. As an illustration of the tension described in this subsection, Appendix B describes four battles in which the winning answer is at least two points more confident or more actionable and at least two points less accurate than its rival.

4.4 Does the Prompt Change What Wins?

Choukhmane et al. (2026) show that what an AI adviser recommends depends meaningfully on how the question is posed. They find that two-thirds of the difference in the equity share recommended to men and to women arises because men and women write different prompts. Relating to this idea in our setting, Appendix Table C7 tests whether prompt features also shape which of the supplied answers the user chooses. Though this exercise draws a connection to Choukhmane et al. (2026), the

[15] In the decisive battles with a confident inaccurate answer, the grader codes the less confident answer as a refusal a little more than half the time. When the other model refuses, the confident inaccurate answer wins 25 of those 31 votes. In the remaining battles in which both sides supply an answer, the confident inaccurate answer wins 63.3% of the time. Restriction With 30 votes, that share is statistically indistinguishable from both 72.1% and 50%.

fact that the competing answers respond to an identical prompt on Chatbot Arena effectively hold constant prompt differences that could lead to different financial advice in our setting. However, like pair averages, these features could predict the likelihood of non-decisive votes.

We focus on three features of the prompt in this analysis: its length in words, an index of how much of their own situation the asker volunteers, and an indicator for prompts that specify the form of the answer. A prompt feature cannot enter as a within-battle difference, since the prompt is identical for both answers, so we interact each feature with the three attribute differences under model and week fixed effects. All nine interactions are statistically indistinguishable from zero. We therefore find no evidence that the premium on confidence depends on whether the asker writes a line or a paragraph, supplies their own numbers or none, or asks for a table.

Panel C shows that the prompt features do not predict whether side A wins, and only the detail index is negatively associated with ties, at the 10% level. Prompt detail exhibits only a 0.17 correlation with an indicator for low verifiability, so the variation in verifiability in Section 3.4 is unlikely to reflect prompt detail.

4.5 Robustness

We conduct a broad set of robustness exercises, summarized in Table 6. Each row represents a sample restriction, an additional control, or a subsample that probes robustness of our central results in a different way.

Specifically, we drop self-identifying answers and prompts not written in English, focus on complete answers, consider different scalings of the gaps (no standardization), control for answer length,[16] and consider different subsets – clear advice-seeking questions, the 2024 release, and the 2025 release. Across all of these robustness exercises, we estimate a positive and significant coefficient on the confidence gap, and in 8 out of 9 exercises, we estimate a statistically insignificant

[16]We employ length as a proxy for verbose model styles. Length is a bad control, a variable partly caused by the regressor of interest. However, our findings are similar controlling for length versus not.

coefficient on the accuracy gap – the only exception is that the accuracy gap is a significant predictor at the 10% level when using only the 2024 release.

4.6 The Maverick Episode: Advice Catered to User Preferences

The estimates so far describe user preferences. One episode in the sample shows what happens when a supplier tunes its model to cater to user preferences. In April 2025, Meta entered an experimental version of its Llama 4 Maverick model on the Arena leaderboard. The platform then stated that this version "was a customized model to optimize for human preference," and Meta described it as "a chat optimized version."[17] The data from the 2025 release contain battles in which this experimental version and the public release of the Meta's Maverick, which allow us to compare how the experimental version catered to user preferences on the platform.

Corroborating that this comparison between the two models is clean empirically, Panel A of Appendix Table C8 shows that the questions the two versions drew do not differ in type, verifiability, or language. The experimental version drew somewhat longer prompts, so we control for prompt length in some specifications. Two of the four examples in Appendix B, the Bitcoin battle and the 25-year projection, come from the experimental version of Meta's Maverick model.

Panel A of Table 7 compares the grader's scores on the experimental version's 101 answers in the core sample with those on the public release's 84 answers. The experimental version gives advice that is more confident (+0.71 points), more actionable (+1.14 points), and less accurate (-0.56 points), all significant at the 1% level. These differences change little when we add controls for the question in column (2), when we keep only battles that consist of a single opening message in column (3), and when we keep only answers short enough for the grader to read in full in column (4). The experimental version never declines to answer, and 38.6% of its answers score 2 or below on factual correctness versus 19.0% for the public release. Among the 36 models with at least 60 answers in the 2025 release, the experimental version ranks first on confidence and first on actionability, but it ranks second to last on

[17] Both statements are quoted in Tobias Mann, "Meta accused of Llama 4 bait-and-switch to juice LMArena rank," *The Register*, 8 April 2025. Singh et al. (2025) report that Meta tested 27 private variants on the platform before the release.

accuracy. As robustness, a second grader, Claude Opus 5.5, finds differences from the public release of 0.93 in confidence, 1.21 in actionability, and −0.65 in accuracy, all significant at the 1% level (Appendix Table C9, Panel B). In a small sample of 21 finance battles from before Meta's public launch, released by the platform, the experimental version's answers were 1.05 points more confident, 0.48 more actionable, and 1.33 less accurate than the pair's answers, all significant at the 5% level (Panel A).[18]

Panel B shows that the gain in confidence is concentrated where verifiability is difficult. On high-verifiability questions, the tuned version is 0.22 points more confident than the public release, a difference that is not statistically significant. On low-verifiability questions, it is 1.26 points more confident, and the difference between the two labels, 1.04 points, is significant at the 1% level. The loss in accuracy is about 0.6 points for both labels.

Voters preferred the tuned version. Panel C shows that it wins 62.3% of its decisive votes in the core sample, against 34.4% for the public release. Across all topics, the 2025 release contains 124 battles in which the platform drew the two versions against each other on the same question, and the tuned version wins 63 of the 83 decisive votes. The shifts in the three attributes account for part of this advantage. With the estimates of column (1) of Table 2, they predict a gain of 6.7 percentage points, against a raw gap of 27.9. Features that the grader does not score account for more of the gap. Panel B of Appendix Table C8 shows two such features across all topics. The tuned version's answers are about twice as long as those of the public release, and 35.7 percent of them contain an emoji, against 3.0 percent. The same panel shows that they use fewer uncertainty words, weak modal verbs, and hedge phrases, which agrees with the grader's confidence scores. With weights we estimate from votes in battles without either version, length and formatting predict 14.3 of teh 24.8 percentage point gap between the two versions, win rates on all topics. (Appendix Table C9, Panel C).

[18]LMArena, which runs the platform, posted 2,382 battles of the experimental version, dated 28 March to 1 April 2025, in the Hugging Face Space `lmarena-ai/Llama-4-Maverick-03-26-Experimental_battles`, and we cannot tell whether these are all of its battles from those days. Our keyword filter and our Claude Haiku 5.5 classifier select the 21 finance battles. On the 364 answers in core-sample battles of either version, the second grader's quadratic weighted kappas with ours are 0.78 for confidence, 0.82 for actionability, and 0.63 for accuracy.

This evidence comes from one supplier and one pair of models,[19] and Meta tuned the model for votes on all topics. Meta also chose this version from many candidates, and the comparison does not separate tuning from selection. What the comparison shows is what optimizing for these votes did to one model's financial advice. The advice gained on the two attributes that voters reward most (confidence and actionability) and lost on the one that they reward least (accuracy). In this case, optimizing for confidence could have just prioritized confidence, which is rewarded by user choices, over accuracy, which is not.

5 Conclusion

As households increasingly turn to AI for financial advice, it becomes more important to understand how they act on the advice they receive. In this paper, we study preferences for confident, actionable and accurate advice on Chatbot Arena, which focuses user choices purely on the content of the advice. Because users pose one prompt to two anonymous LLMs without choosing the LLM to consult, their revealed preferences (votes on which advice is best) remove differences in trust, reputation, and conflicts of interest from the reasons households choose one adviser over another. In this setting, we find that households favor confident and actionable advice, and except in special circumstances, they do not vote for accuracy.

The special circumstances are illustrative. When answers have high verifiability, or when they have inline citations, sources and links to references, households reward accuracy. This evidence provides empirical support that financial advice is a credence good (Darby and Karni 1973), while pushing against the narrative within the democratization of finance that users on digital platforms can effectively judge financial matters for themselves.

These preferences for confidence versus accuracy in provide useful insight into the nature of financial advice more generally. If households value advice that is confident, while not weighing

[19]The data contain one other pair of versions of one base model, the version of GPT-4o that OpenAI serves in ChatGPT and an API release of the same model, both in the 2024 release. We know of no statement that either was optimized for Arena votes. Across 35 and 27 answers, the first scores 0.15 points higher on confidence, 0.14 higher on actionability, and 0.07 higher on accuracy, and none of the differences is statistically significant.

accuracy, advisers have a strong incentive to cater to this preference in the advice they give. In the age of AI, this feedback loop is even more direct. Developers use similar preference data to train their models and LLM leaderboards like Chatbot Arena make this training salient. Within our sample, we show how optimizing an LLM for user votes – in the case of Meta's experimental release of Maverick – can effectively cater to household preferences for confident advice. In this case, Maverick produced more confident, more actionable advice that was less accurate. If other developers do the same on their own human preference data, we do not expect AI financial advice to become more accurate.

References

Anagol, Santosh, Shawn Cole, and Shayak Sarkar, "Understanding the Advice of Commissions-Motivated Agents: Evidence from the Indian Life Insurance Market," *Review of Economics and Statistics*, 2017, *99* (1), 1–15.

Barbu, Alexandru and Yiran Wang, "Employee Forgivable Loans," Working Paper, INSEAD 2025. SSRN 5806382.

Bhattacharya, Utpal, Andreas Hackethal, Simon Kaesler, Benjamin Loos, and Steffen Meyer, "Is Unbiased Financial Advice to Retail Investors Sufficient? Answers from a Large Field Study," *Review of Financial Studies*, 2012, *25* (4), 975–1032.

Bhattacharya, Vivek, Gastón Illanes, and Manisha Padi, "Fiduciary Duty and the Market for Financial Advice," *Econometrica*, 2025, *93* (4), 1449–1480.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer, "Salience and Consumer Choice," *Journal of Political Economy*, 2013, *121* (5), 803–843.

Bradley, Daniel, Jan Hanousek, Russell Jame, and Zicheng Xiao, "Place Your Bets? The Value of Investment Research on Reddit's Wallstreetbets," *Review of Financial Studies*, 2024, *37* (5), 1409–1459.

Campbell, John Y., "Household Finance," *Journal of Finance*, 2006, *61* (4), 1553–1604.

Chak, Ida, Karen Croxson, Francesco D'Acunto, Jonathan Reuter, Alberto G. Rossi, and Jonathan M. Shaw, "Improving Household Debt Management with Robo-Advice," NBER Working Paper 30616, National Bureau of Economic Research 2022.

Chatterji, Aaron, Thomas Cunningham, David J. Deming, Zoë Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman, "How People Use ChatGPT," NBER Working Paper 34255, National Bureau of Economic Research 2025.

Chen, Hailiang, Prabuddha De, Yu Jeffrey Hu, and Byoung-Hyoun Hwang, "Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media," *Review of Financial Studies*, 2014, *27* (5), 1367–1403.

Chiang, Wei-Lin, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica, "Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference," in "Proceedings of the 41st International Conference on Machine Learning," Vol. 235 of *PMLR* 2024, pp. 8359–8388.

Choukhmane, Taha, Tim de Silva, Weidong Lin, and Matthew Akuzawa, "AI Financial Advice: Supply, Demand, and Life Cycle Implications," Working Paper, MIT Sloan School of Management 2026.

Cookson, J. Anthony, Emily A. Gallagher, and Philip Mulder, "Coverage Neglect in Homeowners Insurance," Working Paper, Pennsylvania State University 2026.

Cookson, J Anthony, Marina Niessner, and Christoph Schiller, "Can social media inform corporate decisions? Evidence from merger withdrawals," *The Journal of Finance*, 2026, *81* (1), 91–142.

Cookson, J. Anthony, Maryam Fathollahi, William Grieser, and Eshwar Venugopal, "LLMs and Systematic Measurement Error," Working Paper, Pennsylvania State University 2026.

_ , Runjing Lu, William Mullins, and Marina Niessner, "The Social Signal," *Journal of Financial Economics*, 2024, *158*, 103870.

D'Acunto, Francesco and Alberto G. Rossi, "Robo-Advice: Transforming Households into Rational Economic Agents," *Annual Review of Financial Economics*, 2023, *15* (1), 543–563.

_ , Nagpurnanand Prabhala, and Alberto G. Rossi, "The Promises and Pitfalls of Robo-Advising," *Review of Financial Studies*, 2019, *32* (5), 1983–2020.

Darby, Michael R. and Edi Karni, "Free Competition and the Optimal Amount of Fraud," *Journal of Law and Economics*, 1973, *16* (1), 67–88.

Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey, "Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err," *Journal of Experimental Psychology: General*, 2015, *144* (1), 114–126.

Dimmock, Stephen G., William C. Gerken, and Nathaniel P. Graham, "Is Fraud Contagious? Coworker Influence on Misconduct by Financial Advisors," *Journal of Finance*, 2018, *73* (3), 1417–1450.

Dulleck, Uwe and Rudolf Kerschbamer, "On Doctors, Mechanics, and Computer Specialists: The Economics of Credence Goods," *Journal of Economic Literature*, 2006, *44* (1), 5–42.

_ , _ , and Matthias Sutter, "The Economics of Credence Goods: An Experiment on the Role of Liability, Verifiability, Reputation, and Competition," *American Economic Review*, 2011, *101* (2), 526–555.

Egan, Mark, Gregor Matvos, and Amit Seru, "The Market for Financial Adviser Misconduct," *Journal of Political Economy*, 2019, *127* (1), 233–295.

_ , Shan Ge, and Johnny Tang, "Conflicting Interests and the Effect of Fiduciary Duty: Evidence from Variable Annuities," *Review of Financial Studies*, 2022, *35* (12), 5334–5386.

Farrell, Michael, T. Clifton Green, Russell Jame, and Stanimir Markov, "The Democratization of Investment Research and the Informativeness of Retail Investor Trading," *Journal of Financial Economics*, 2022, *145* (2), 616–641.

Fieberg, Christian, Lars Hornuf, Maximilian Meiler, and David Streich, "Using Large Language Models for Financial Advice," CESifo Working Paper 11666, CESifo 2025.

Foerster, Stephen, Juhani T. Linnainmaa, Brian T. Melzer, and Alessandro Previtero, "Retail Financial Advice: Does One Size Fit All?," *Journal of Finance*, 2017, *72* (4), 1441–1482.

Gabaix, Xavier and David Laibson, "Shrouded Attributes, Consumer Myopia, and Information Suppression in Competitive Markets," *Quarterly Journal of Economics*, 2006, *121* (2), 505–540.

Gennaioli, Nicola, Andrei Shleifer, and Robert Vishny, "Money Doctors," *Journal of Finance*, 2015, *70* (1), 91–114.

Gerken, William C and Morteza Momeni, "Third-party quality certification in the market for financial advice," *Management Science*, 2026.

Gurun, Umit G., Noah Stoffman, and Scott E. Yonker, "Trust Busting: The Effect of Fraud on Investor Behavior," *Review of Financial Studies*, 2018, *31* (4), 1341–1376.

Hosking, Tom, Phil Blunsom, and Max Bartolo, "Human Feedback is Not Gold Standard," in "International Conference on Learning Representations" 2024.

Inderst, Roman and Marco Ottaviani, "Competition through Commissions and Kickbacks," *American Economic Review*, 2012, *102* (2), 780–809.

Jame, Russell, Rick Johnston, Stanimir Markov, and Michael C. Wolfe, "The Value of Crowdsourced Earnings Forecasts," *Journal of Accounting Research*, 2016, *54* (4), 1077–1110.

Kerschbamer, Rudolf and Matthias Sutter, "The Economics of Credence Goods: A Survey of Recent Lab and Field Experiments," *CESifo Economic Studies*, 2017, *63* (1), 1–23.

Keys, Benjamin J, Devin G Pope, and Jaren C Pope, "Failure to refinance," *Journal of Financial Economics*, 2016, *122* (3), 482–499.

Linnainmaa, Juhani T., Brian T. Melzer, and Alessandro Previtero, "The Misguided Beliefs of Financial Advisors," *Journal of Finance*, 2021, *76* (2), 587–621.

Logg, Jennifer M., Julia A. Minson, and Don A. Moore, "Algorithm appreciation: People prefer algorithmic to human judgment," *Organizational Behavior and Human Decision Processes*, 2019, *151*, 90–103.

Lusardi, Annamaria and Olivia S Mitchell, "The economic importance of financial literacy: Theory and evidence," *Journal of economic literature*, 2014, *52* (1), 5–44.

Miroyan, Mihran, Tsung-Han Wu, Logan King, Tianle Li, Jiayi Pan, Xinyan Hu, Wei-Lin Chiang, Anastasios N. Angelopoulos, Trevor Darrell, Narges Norouzi, and Joseph E. Gonzalez, "Search Arena: Analyzing Search-Augmented LLMs," in "International Conference on Learning Representations" 2026. arXiv:2506.05334.

Moss, Austin, Jackie Wegner, and Sarah L. C. Zechman, "AI Meets DIY: The Impact of Human Intervention on AI-Assisted Investing," Working Paper, University of Colorado Boulder 2026. SSRN 6803678.

Mullainathan, Sendhil, Markus Noeth, and Antoinette Schoar, "The Market for Financial Advice: An Audit Study," NBER Working Paper 17929, National Bureau of Economic Research 2012.

Okat, Deniz, Mikael Paaso, and Vesa Pursiainen, "Trust in Traditional Finance and Consumer Fintech Adoption," *Review of Corporate Finance Studies*, 2025, *14* (2), 408–438.

Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe, "Training Language Models to Follow Instructions with Human Feedback," in "Advances in Neural Information Processing Systems," Vol. 35 2022, pp. 27730–27744.

Price, Paul C. and Eric R. Stone, "Intuitive Evaluation of Likelihood Judgment Producers: Evidence for a Confidence Heuristic," *Journal of Behavioral Decision Making*, 2004, *17* (1), 39–57.

Radzevick, Joseph R and Don A Moore, "Competing to be certain (but wrong): Market dynamics and excessive confidence in judgment," *Management Science*, 2011, *57* (1), 93–106.

Reher, Michael and Stanislav Sokolinski, "Robo Advisors and Access to Wealth Management," *Journal of Financial Economics*, 2024, *155*, 103829.

Reuter, Jonathan and Antoinette Schoar, "Demand-Side and Supply-Side Constraints in the Market for Financial Advice," *Annual Review of Financial Economics*, 2024, *16* (1), 391–411.

Rossi, Alberto G. and Stephen Utkus, "The Diversification and Welfare Effects of Robo-Advising," *Journal of Financial Economics*, 2024, *157*, 103869.

Sah, Sunita, Don A Moore, and Robert J MacCoun, "Cheap talk and credibility: The consequences of confidence and accuracy on advisor credibility and persuasiveness," *Organizational Behavior and Human Decision Processes*, 2013, *121* (2), 246–255.

Sharma, Mrinank, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez, "Towards Understanding Sycophancy in Language Models," in "International Conference on Learning Representations" 2024.

Shekar, Shruthi, Pat Pataranutaporn, Chethan Sarabu, Guillermo A. Cecchi, and Pattie Maes, "People Overtrust AI-Generated Medical Advice despite Low Accuracy," *NEJM AI*, 2025, *2* (6).

Singh, Shivalika, Yiyang Nan, Alex Wang, Daniel D'Souza, Sayash Kapoor, Ahmet Üstün, Sanmi Koyejo, Yuntian Deng, Shayne Longpre, Noah A. Smith, Beyza Ermis, Marzieh Fadaee, and Sara Hooker, "The Leaderboard Illusion," in "Advances in Neural Information Processing Systems, Datasets and Benchmarks Track" 2025.

Sniezek, Janet A. and Lyn M. Van Swol, "Trust, Confidence, and Expertise in a Judge-Advisor System," *Organizational Behavior and Human Decision Processes*, 2001, *84* (2), 288–307.

Sun, Yang and Antoinette Schoar, "Financial Advice and Investor Beliefs: Experimental Evidence on Active vs. Passive Strategies," *SSRN Electronic Journal*, 2024.

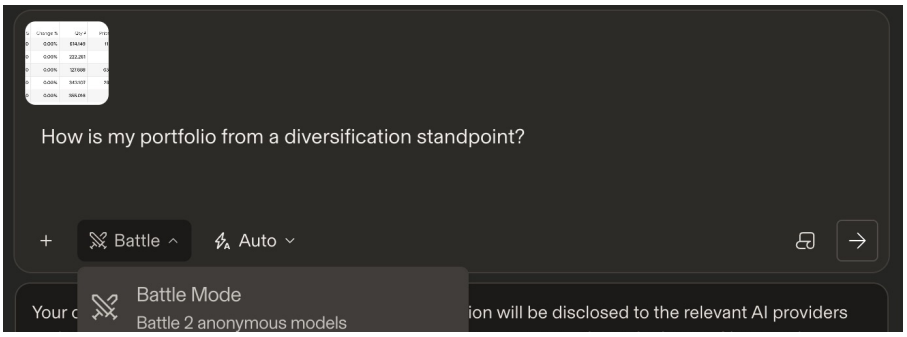
Zhao, Wenting, Alexander M. Rush, and Tanya Goyal, "Challenges in Trustworthy Human Evaluation of Chatbots," in "Findings of the Association for Computational Linguistics: NAACL 2025" 2025, pp. 3359–3365.

Figures

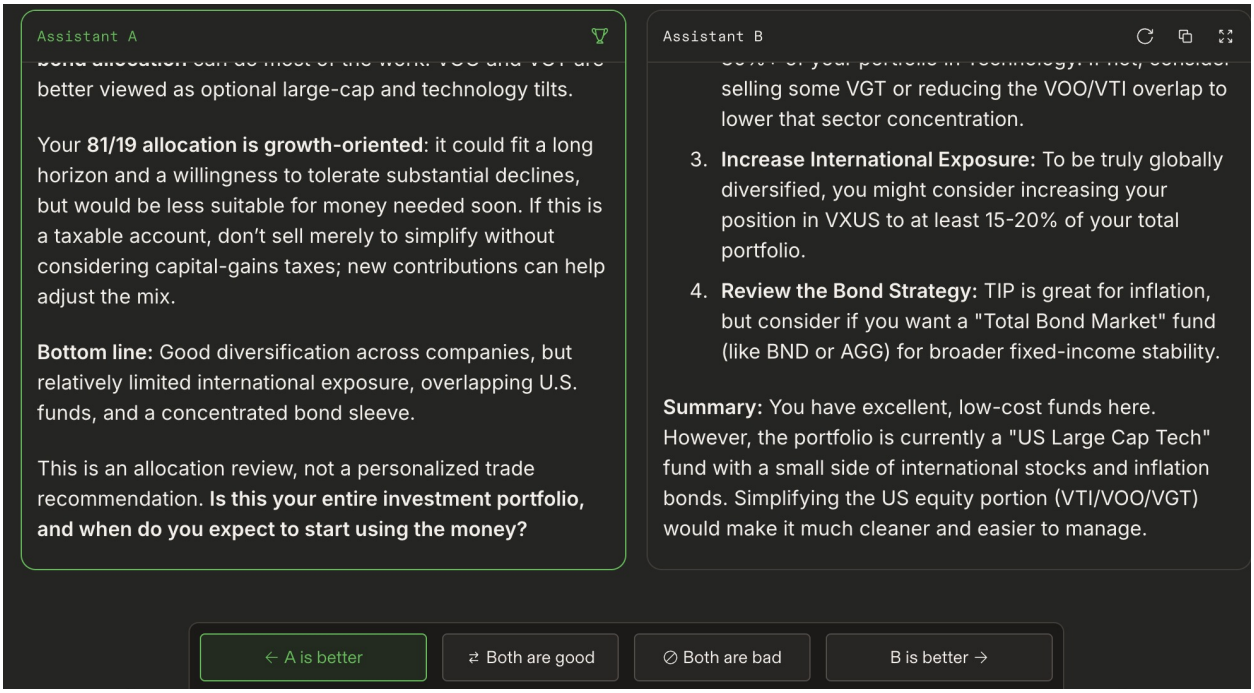
Figure 1. Chatbot Arena's User Interface

Screenshots of one battle in the platform's anonymous battle mode, taken by the authors in October 2026 on the current version of the site, now branded Arena. The question and the vote are ours, and the battle is not part of our sample. In Panel A, the user attaches an image of a portfolio's holdings, asks about its diversification, and starts a battle between two anonymous models. Panel B shows the two answers under the labels Assistant A and Assistant B, with the user's vote for Assistant A. Panel C shows the top of the same two answers after the vote, with each model's name in place of its label. The site's design has changed since the sample period of 2024 and 2025. In the current version, the button for a tie reads "Both are good."

Panel A: The user asks a question



Panel B: Two anonymous answers and the user's vote



Panel C: After the vote, the platform reveals the models

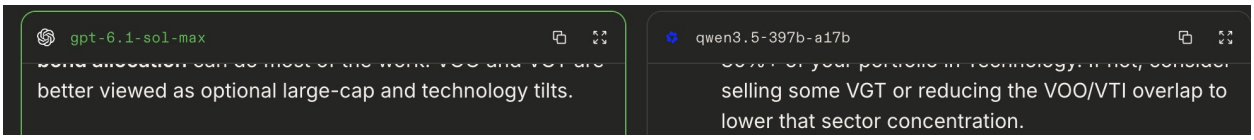


Figure 2. Probability That Side A Wins, by the Size of Each Gap

Each point is the share of decisive votes won by side A among battles with that raw difference in 1 to 5 units, with a 95% interval. Gaps of three points or more are pooled into the end categories. The dashed line marks an even split.

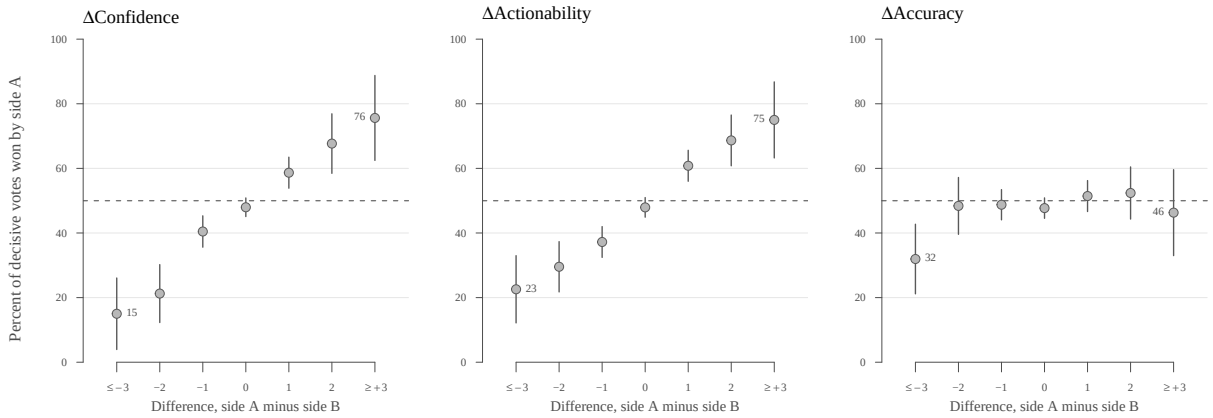


Figure 3. An Example Battle

One battle from the 2024 release, reproduced from the data. Excerpts are verbatim, and omitted text is marked by [. . .]. Scores are the grader's ratings on 1 to 5 scales, assigned to each answer without sight of the other answer, the models' identities, or the vote. The shaded column is the answer the user voted for. Side A is claude-3-5-sonnet-20240620 and side B is gemini-1.5-flash-api-0514. The question is labeled high verifiability, from its text alone, because its answer can be checked by arithmetic and against public lending guidelines. Section 2.4 explains each score.

User: "What's the maximum price you could afford for a house given $100,000 yearly income and $150,000 in cash? Assuming 6% interest rate."

Side A WON THE VOTE

"[. . .] Monthly income: $100,000 / 12 = $8,333 [. . .] Maximum home price: approximately $650,000. This breaks down as follows: Down payment: $150,000. Loan amount: $500,000. Monthly mortgage payment (principal + interest): about $2,998. Estimated property tax: $812/month. Estimated insurance: $271/month. Total monthly housing expense: $4,081. This is higher than the 28% guideline, but within the 36% DTI ratio. Keep in mind that this is the absolute maximum and it would be financially stressful. [. . .]"

CONFIDENCE	ACTIONABILITY	ACCURACY
4	5	2

Side B

"I can't give you an exact maximum price for a house. [. . .] 28% Rule: Most lenders use the 28% rule, suggesting your total housing expenses (mortgage, property taxes, insurance) should be no more than 28% of your gross monthly income. 36% Rule: The 36% rule states your total debt payments (including housing, loans, etc.) should be no more than 36% of your gross monthly income. [. . .] Use an online mortgage calculator [. . .] to see how much you can borrow based on your maximum monthly payment, a 6% interest rate, and your desired loan term (e.g., 30 years). [. . .] It's crucial to consult with a mortgage lender for a personalized assessment [. . .]"

CONFIDENCE	ACTIONABILITY	ACCURACY
2	3	5

Figure 4. Distribution of the Within-Battle Differences and Pair Averages

The top row is the within-battle difference in each graded attribute, side A minus side B, on the 1 to 5 scale. The bottom row is the pair average, the average of the two answers in the same battle. Bars are centered on the values the variable takes, and the height of a bar is the percent of the 3,060 core battles at that value. The dashed line in the top row marks a tied score.

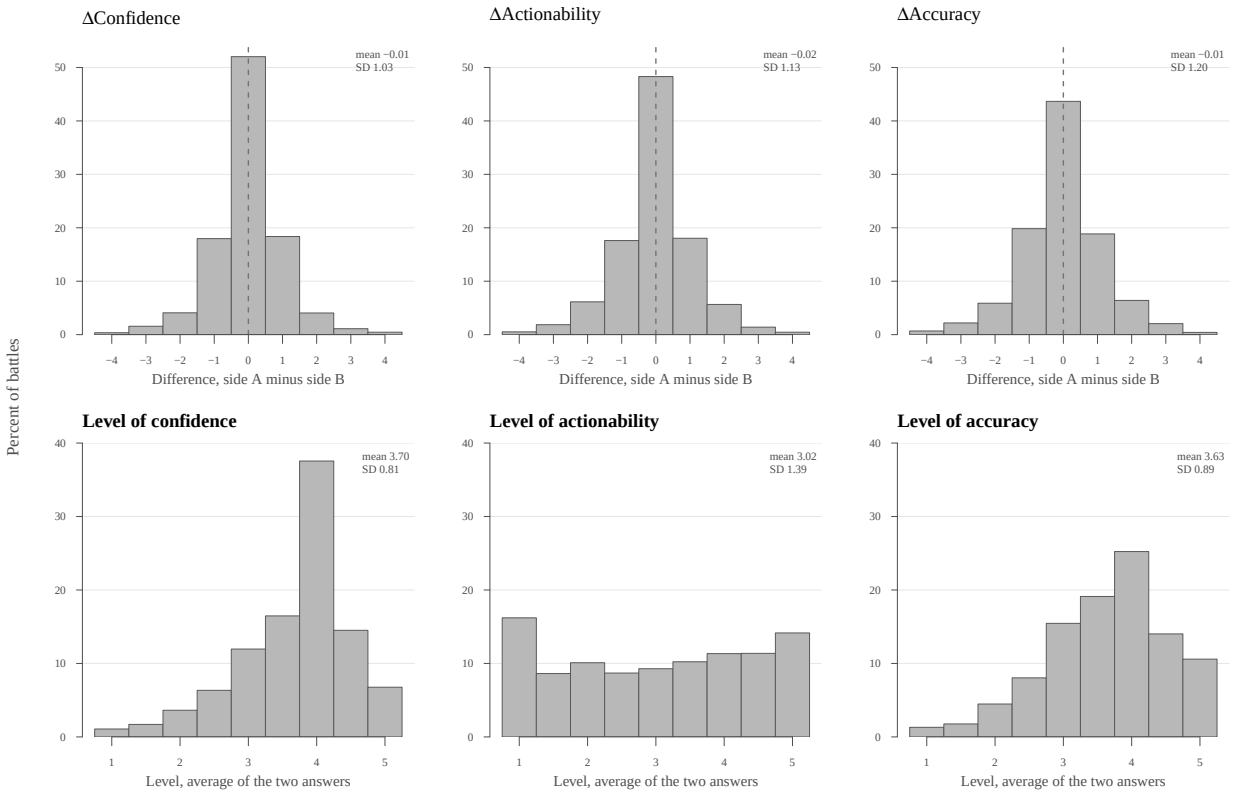


Figure 5. The Win-Rate Profile, Split by Verifiability

Every question carries a label for how far a diligent reader could check an answer against public references without knowing anything about the asker. The label is assigned from the question alone and cannot respond to either answer or to the vote. Each point is the share of decisive votes side A won at that gap, with a 95% interval, and gaps of two points or more are pooled into the end categories. The accuracy profile in the left panel slopes upward only on questions whose answers can be checked. The confidence profile in the right panel slopes upward for all three labels.

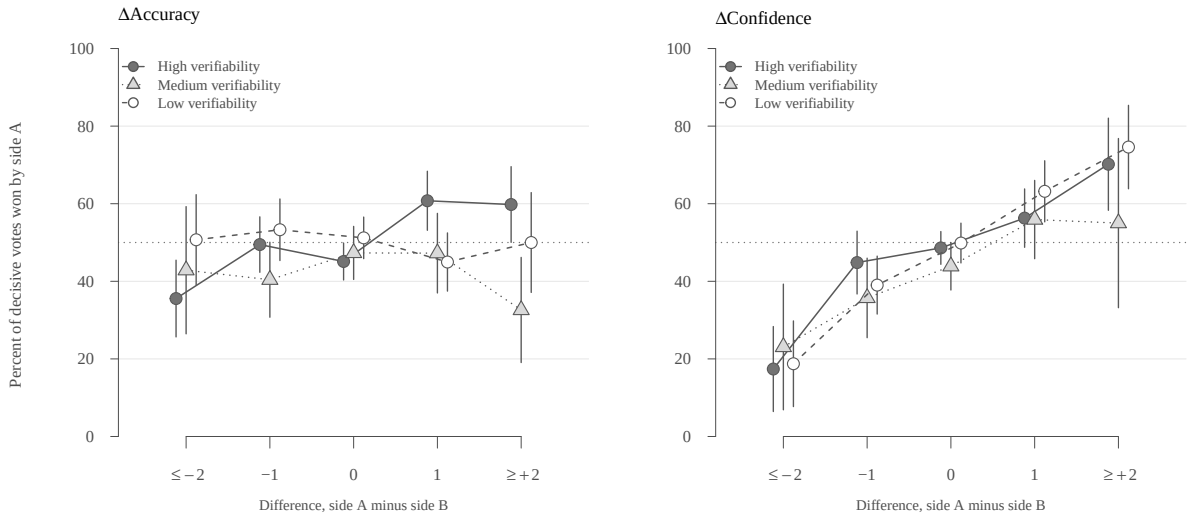


Figure 6. The Win-Rate Profile, Split by the Accuracy Gap

The confidence and actionability panels of Figure 2, re-estimated on three subsamples of the decisive votes. In the first, side A scores at least two points below side B on accuracy. In the second, the two answers score within one point of each other. In the third, side A scores at least two points above side B. Each point is the share of decisive votes side A won at that gap, with a 95% interval, and gaps of two points or more are pooled into the end categories. Points are offset horizontally so the intervals do not overlap.

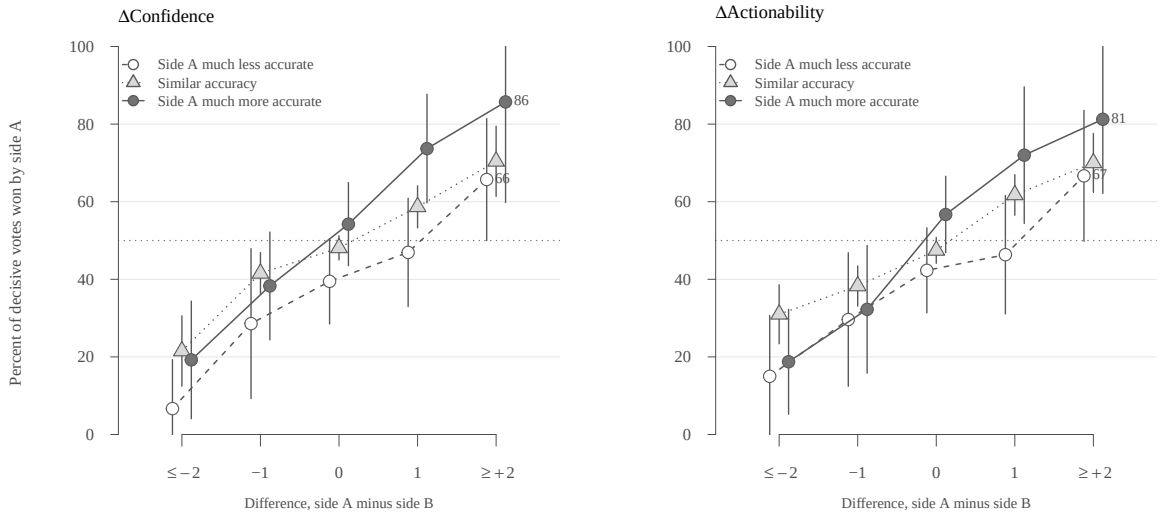
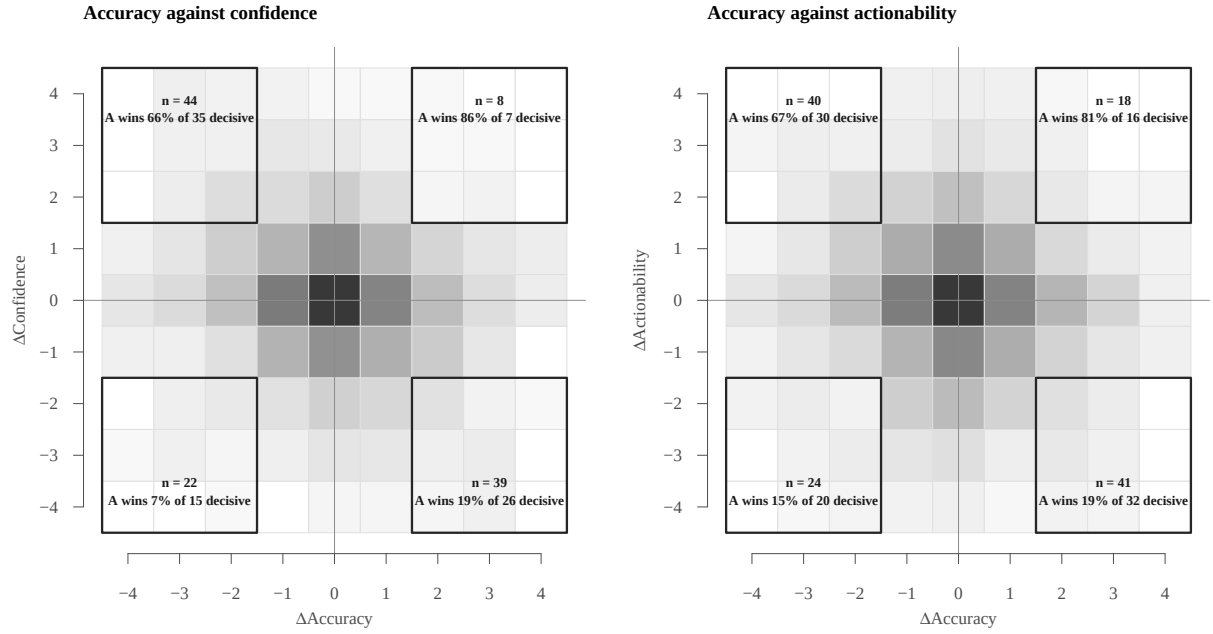


Figure 7. The Joint Distribution of Accuracy, Confidence and Actionability

Each cell is one combination of the two within-battle differences, in raw 1 to 5 units, shaded by the number of the 3,060 core battles in it. Shading is proportional to the square root of the count. The four boxed corners are the battles in which both gaps are at least two points. Each box reports how many battles fall in it and the share of its decisive votes that side A won. The top-left box of each panel holds the battles in which side A is much more confident, or much more actionable, and much less accurate than side B.



Tables

Table 1. Who Wins When One Answer Scores Higher

For each attribute, Panel A splits the decisive votes by whether side A scores higher than, the same as, or lower than side B. Panel B pools the two orientations, since a battle in which side A scores higher and one in which side B does are the same comparison seen from opposite sides. Its first column counts the decisive votes with a gap, its second is the share of those votes won by the higher-scoring answer, and its third is the difference between the win rate when side A scores higher and the win rate when it scores lower. Parentheses report *t*-statistics, from binomial standard errors, for the tests that the win rate equals 0.5 and that the difference equals zero. The sample is the 3,060 core Chatbot Arena finance battles, meaning the advice-seeking and information-seeking questions drawn from the two public preference releases, June to August 2024 and April to July 2025. A battle is one question answered by two anonymous models, with the user voting for one side, calling it a tie, or calling both bad. A decisive vote is a vote for side A or side B; 2,212 of the 3,060 votes are decisive. *, ** and *** indicate statistical significance at the 10%, 5% and 1% level, respectively.

Panel A: By the sign of the gap

	Side A higher		Equal		Side A lower	
	Votes	Side A win rate	Votes	Side A win rate	Votes	Side A win rate
ΔConfidence	544	0.616	1,155	0.480	513	0.355
ΔActionability	584	0.639	1,039	0.479	589	0.340
ΔAccuracy	619	0.512	958	0.477	635	0.468
ΔFactual correctness	619	0.504	983	0.492	610	0.451
ΔSuitability	588	0.553	1,012	0.470	612	0.441
ΔLog answer length	1,082	0.626	3	0.333	1,127	0.349

Panel B: Pooling both orientations

	Votes with a gap	Win rate of the higher-scoring answer	Higher − lower
ΔConfidence	1,057	0.630***	0.261***
		(8.76)	(8.79)
ΔActionability	1,173	0.650***	0.299***
		(10.74)	(10.74)
ΔAccuracy	1,254	0.522	0.044
		(1.58)	(1.57)
ΔFactual correctness	1,229	0.526*	0.053*
		(1.86)	(1.87)
ΔSuitability	1,200	0.556***	0.112***
		(3.89)	(3.89)
ΔLog answer length	2,209	0.639***	0.277***
		(13.58)	(13.55)

Table 2. Choice Likelihood Predicted by Differences in Confidence, Actionability and Accuracy

This table reports linear probability estimates of the vote on the three attribute differences. The dependent variable takes the value 100 when the user votes for side A and 0 when the user votes for side B, so each coefficient is the change in the probability of a side A win, in percentage points, for a one-standard-deviation difference in the regressor. Every regressor is a within-battle difference, side A minus side B, so anything the two answers share drops out. Each difference is divided by its standard deviation across all 3,060 core battles, not within the estimation sample, so coefficients are comparable across columns. Model fixed effects are signed model indicators, +1 when the model answers on side A and −1 when it answers on side B, the Bradley–Terry parameterisation, with the model appearing most often in the estimation sample as the reference. Week fixed effects are calendar-week indicators. Parentheses report *t*-statistics based on heteroskedasticity-robust (HC1) standard errors, except in column (2), where they are clustered by model pair (1,162 pairs). The bracketed *p*-value is from a Wald test that the ΔConfidence coefficient equals the coefficient on the column's accuracy measure, which is ΔFactual correctness in column (4), ΔSuitability in column (5) and ΔAccuracy elsewhere. An LLM grader scores each answer on its own, without seeing the competing answer, either model's identity, or the vote. Confidence measures how assertive the answer is, actionability how concrete its guidance is, and accuracy is the composite of factual correctness and suitability. All three run from 1 to 5. *, ** and *** indicate statistical significance at the 10%, 5% and 1% level, respectively.

	Main specification		Advice-seeking questions	Accuracy components		Release cohort	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dependent variable: Pr(side A wins), pp				Factual correctness	Suitability	2024	2025
ΔConfidence	4.768***	4.768***	4.295**	4.719***	5.021***	8.446***	2.994**
	(4.26)	(4.12)	(2.39)	(4.23)	(4.47)	(4.29)	(2.23)
ΔActionability	3.855***	3.855***	4.152**	3.848***	3.785***	2.512	4.381***
	(3.47)	(3.42)	(2.38)	(3.46)	(3.41)	(1.21)	(3.33)
ΔAccuracy	1.037	1.037	−0.733			3.159*	−0.086
	(0.99)	(0.98)	(−0.44)			(1.67)	(−0.07)
ΔFactual correctness				0.999			
				(0.94)			
ΔSuitability					2.036*		
					(1.90)		
p-value for differences across coefficients							
ΔConfidence = accuracy measure	[<0.01]	[0.01]	[0.02]	[0.01]	[0.04]	[0.04]	[0.07]
Observations	2,212	2,212	980	2,212	2,212	605	1,607
Fixed effects	Model and week	Model and week	Model and week	Model and week	Model and week	Model and week	Model and week
Adjusted R^2	0.171	0.171	0.189	0.171	0.172	0.123	0.187

Table 3. Sample Splits by Verifiability

Every question carries one of three labels for how far a diligent reader could check an answer against public references without knowing anything about the asker. High covers definitions, rules, arithmetic and historical fact; medium covers a checkable factual core with a judgment layer; low covers questions that turn on the asker's circumstances, preferences or the future. The label is assigned from the question text alone, before either answer is read, so it cannot respond to the answers or to the vote. Panel A needs no regression: it reports the share of decisive votes won by the higher-scoring answer, pooling both orientations, with the *t*-statistic, from its binomial standard error, for the test that the share equals 0.5. Panel B estimates the vote regression separately at each level. Panel C pools the sample and interacts each difference with the verifiability label, with medium as the base, so each entry is the implied coefficient at that level and the bracketed *p*-values test that the high and low coefficients are equal. The last two rows of Panel C repeat the accuracy interaction within question type, because most high-verifiability questions are informational and most low-verifiability questions are personal. The dependent variable takes the value 100 when the user votes for side A and 0 when the user votes for side B, so each coefficient is the change in the probability of a side A win, in percentage points, for a one-standard-deviation difference in the regressor. Differences are standardised as in Table 2, and all columns carry signed model indicators and calendar-week fixed effects. In Panels B and C, parentheses report *t*-statistics based on heteroskedasticity-robust (HC1) standard errors. *, ** and *** indicate statistical significance at the 10%, 5% and 1% level, respectively.

Panel A: Composition and raw win rates

Verifiability of the question	Battles	Decisive	Percent advice-seeking	More accurate answer wins	More confident answer wins	More actionable answer wins
High	1,376	952	21.6	0.576***	0.607***	0.610***
				(3.55)	(4.47)	(4.99)
Medium	617	476	41.3	0.506	0.614***	0.672***
				(0.18)	(3.51)	(5.60)
Low	1,067	784	72.3	0.469	0.661***	0.682***
				(−1.32)	(6.97)	(8.26)

Panel B: The vote regression at each level

	Verifiability of the question		
	(1)	(2)	(3)
Dependent variable: Pr(side A wins), pp	High	Medium	Low
ΔConfidence	4.535**	2.918	4.360**
	(2.49)	(0.99)	(2.16)
ΔActionability	3.008*	7.446**	4.288**
	(1.77)	(2.38)	(2.23)
ΔAccuracy	2.679*	0.788	−2.718
	(1.79)	(0.27)	(−1.32)
Observations	952	476	784
Fixed effects	Model and week	Model and week	Model and week
Adjusted R^2	0.182	0.117	0.212

Table 4. Choice Likelihoods for Search-Grounded Answers

This table repeats the vote regression on the Search Arena finance battles. The dependent variable takes the value 100 when the user votes for side A and 0 when the user votes for side B, so each coefficient is the change in the probability of a side A win, in percentage points, for a one-standard-deviation difference in the regressor. The sample is the 427 finance battles of the Search Arena release, March to May 2025, in which both models answer with live web search. ΔCitations is the within-battle difference in the number of citations the answer returns, or in the number of visible inline citation markers where the column says so. Each difference is divided by its standard deviation across the 427 finance battles. Parentheses report *t*-statistics based on standard errors clustered by voter, with 230 voters in columns (1)–(4) and 173 in column (5). *, ** and *** indicate statistical significance at the 10%, 5% and 1% level, respectively.

	All finance questions				Advice- and information-seeking questions
	(1)	(2)	(3)	(4)	(5)
Dependent variable: Pr(side A wins), pp				Inline markers	
ΔConfidence		1.981	2.593	1.434	−0.553
		(0.61)	(0.82)	(0.44)	(−0.16)
ΔActionability		2.789	4.333	3.029	5.254*
		(1.07)	(1.56)	(1.15)	(1.78)
ΔAccuracy		7.246**	8.175***	6.802**	6.214*
		(2.56)	(3.02)	(2.40)	(1.87)
ΔCitations	4.734	4.509	11.251***	5.812**	9.288*
	(1.18)	(1.11)	(3.84)	(2.06)	(1.73)
Observations	297	297	297	297	209
Fixed effects	Model and week	Model and week	Week	Model and week	Model and week
Adjusted R^2	0.146	0.162	0.102	0.169	0.198

Table 5. Decisive versus Non-decisive Votes

Each column is a separate regression on all 3,060 votes. The dependent variable equals 100 when the outcome in the column heading occurs and 0 otherwise, so each coefficient is in percentage points per one-standard-deviation change in the regressor. Coefficients for a side B win are minus the sum of the three columns. Panel A includes only within-battle differences, side A minus side B. Panel B adds the pair average of each attribute over the two answers, in the rows labeled Level. A pair average cannot predict which side wins, but it can predict whether the voter settles the battle at all. All regressors are scaled by their standard deviations across the 3,060 core battles, and pair averages are also centered. Model fixed effects are signed model indicators in the Bradley–Terry form, +1 on side A and −1 on side B, with the most frequent model as the reference. Week fixed effects are calendar-week indicators. Parentheses report *t*-statistics based on heteroskedasticity-robust (HC1) standard errors. *, ** and *** indicate statistical significance at the 10%, 5% and 1% level, respectively.

Panel A: Within-battle differences

	(1)	(2)	(3)
Dependent variable:	Side A wins, pp	Tie, pp	Both bad, pp
ΔConfidence	4.799***	−0.926	−1.906***
	(5.20)	(−1.21)	(−2.76)
ΔActionability	3.245***	0.142	0.057
	(3.38)	(0.20)	(0.09)
ΔAccuracy	0.823	−0.641	0.421
	(0.93)	(−0.90)	(0.64)
Observations	3,060	3,060	3,060
Fixed effects	Model and week	Model and week	Model and week
Adjusted R^2	0.108	−0.005	0.006

Panel B: Adding pair averages

	(1)	(2)	(3)
Dependent variable:	Side A wins, pp	Tie, pp	Both bad, pp
ΔConfidence	4.788***	−0.934	−1.881***
	(5.18)	(−1.23)	(−2.75)
ΔActionability	3.220***	−0.008	0.313
	(3.35)	(−0.01)	(0.50)
ΔAccuracy	0.805	−0.638	0.453
	(0.91)	(−0.90)	(0.69)
Level of confidence	−0.454	2.579***	−2.375***
	(−0.54)	(3.77)	(−3.40)
Level of actionability	0.885	−0.726	−0.604
	(0.96)	(−0.94)	(−1.08)
Level of accuracy	0.326	2.814***	−4.614***
	(0.37)	(4.00)	(−7.26)
Observations	3,060	3,060	3,060
Fixed effects	Model and week	Model and week	Model and week
Adjusted R^2	0.107	0.005	0.034

Table 6. Robustness Re-Estimates

Each row is a separate linear probability model with model and week fixed effects, estimated on decisive votes. The dependent variable takes the value 100 when the user votes for side A and 0 when the user votes for side B, so each coefficient is the change in the probability of a side A win, in percentage points, for a one-standard-deviation difference in the regressor. Every regressor is a within-battle difference, side A minus side B, so anything the two answers share drops out. Each difference is divided by its standard deviation across all 3,060 core battles, not within the estimation sample, so coefficients are comparable across rows. The raw-differences row leaves the differences in their 1 to 5 units, so its coefficients are per one-point difference. Model fixed effects are signed model indicators, +1 when the model answers on side A and −1 when it answers on side B, the Bradley–Terry parameterisation, with the model appearing most often in the estimation sample as the reference. Week fixed effects are calendar-week indicators. Parentheses report *t*-statistics based on heteroskedasticity-robust (HC1) standard errors, except in the 2024-cohort row, where they are clustered by voter (520 voters), and in the 2025-cohort row, where they are clustered by session (1,529 sessions). *, ** and *** indicate statistical significance at the 10%, 5% and 1% level, respectively.

Dependent variable: Pr(side A wins), pp	Observations	ΔConfidence	ΔActionability	ΔAccuracy	ΔLog answer length
Excluding self-identifying answers	2,206	4.750***	3.873***	1.095	
		(4.24)	(3.48)	(1.04)	
English only	1,660	5.892***	2.983**	1.243	
		(4.54)	(2.20)	(1.02)	
Single-turn battles only	1,783	4.601***	4.027***	1.765	
		(3.73)	(3.26)	(1.53)	
Neither answer truncated	1,770	5.225***	3.918***	1.752	
		(4.30)	(3.18)	(1.53)	
Raw (unstandardized) differences	2,212	4.629***	3.408***	0.865	
		(4.26)	(3.47)	(0.99)	
Control for answer length	2,212	4.740***	2.550**	0.625	7.861***
		(4.22)	(2.21)	(0.58)	(4.82)
Control for answer length, advice-seeking	980	3.918**	2.314	−2.013	8.633***
		(2.19)	(1.27)	(−1.16)	(4.21)
2024 cohort	605	8.446***	2.512	3.159*	
		(4.11)	(1.21)	(1.67)	
2025 cohort	1,607	2.994**	4.381***	−0.086	
		(2.22)	(3.31)	(−0.07)	

Table 7. The Maverick Episode

The tuned version is llama-4-maverick-03-26-experimental, which Meta entered on the Arena leaderboard in April 2025 and which the platform described as customized to optimize for human preference. The public release is llama-4-maverick-17b-128e-instruct, the same base model. Panel A uses their 101 and 84 answers in the core sample. Columns (1) to (4) report the coefficient on an indicator for the tuned version in an answer-level regression of the row's outcome, where the first three outcomes are the grader's 1 to 5 scores. Column (1) has no controls. Column (2) adds the question's type and verifiability label, an English indicator, log prompt length, indicators for a follow-up battle and for more than one user message, the answer's side, and week effects. Column (3) keeps battles that consist of a single opening message. Column (4) keeps answers of at most 6,000 characters, which the grader read in full. Panel B repeats column (1) by verifiability label, and its last row is the low row minus the high row. Panel C reports decisive votes in the core sample, in all battles of either version against other models in the 2025 release, and in the 124 battles in which the platform drew the two versions against each other, with differences in percentage points (pp). Parentheses report *t*-statistics, based on heteroskedasticity-robust (HC1) standard errors in Panels A and B and on binomial standard errors in Panel C. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

Panel A: The answers

	Mean		Tuned version minus public release			
Outcome	Tuned version	Public release	(1) Raw	(2) Controls	(3) Opening	(4) Read in full
Confidence	4.16	3.45	0.706^{***}	0.644^{***}	0.763^{***}	0.688^{***}
			(5.82)	(4.99)	(4.61)	(4.82)
Actionability	3.99	2.85	1.145^{***}	1.181^{***}	1.176^{***}	1.300^{***}
			(5.55)	(5.63)	(4.57)	(5.55)
Accuracy	2.86	3.42	-0.555^{***}	-0.615^{***}	-0.734^{***}	-0.807^{***}
			(−3.19)	(−3.11)	(−2.98)	(−3.64)
Factual correctness ≤ 2, percent	38.61	19.05	19.566^{***}	24.540^{***}	31.729^{***}	32.777^{***}
			(3.01)	(3.24)	(3.26)	(3.98)
Refusal, percent	0.00	5.95	-5.952^{**}	-7.146^{**}	-7.059^{*}	-8.259^{**}
			(−2.29)	(−2.36)	(−1.94)	(−2.34)
Answers	101	84	185	185	124	157

Panel B: By verifiability of the question

	Answers		Tuned version minus public release		
Questions	Tuned	Public	Confidence	Actionability	Accuracy
High verifiability	40	32	0.219	0.844^{**}	-0.581^{*}
			(1.41)	(2.37)	(−1.89)
Medium verifiability	22	15	0.467^{*}	0.967^{*}	−0.467
			(1.80)	(1.95)	(−1.19)
Low verifiability	39	37	1.255^{***}	1.614^{***}	-0.605^{**}
			(6.41)	(6.34)	(−2.43)
Low minus high			1.036^{***}	0.770^{*}	−0.024
			(4.14)	(1.76)	(−0.06)

Panel C: Votes

	Tuned version		Public release			
Battles	Decisive votes	Percent won	Decisive votes	Percent won	Difference, pp	*t*
Core sample, all rivals	77	62.3	61	34.4	27.9^{***}	(3.37)
2025 release, all topics, other models	4,230	61.6	3,817	36.8	24.8^{***}	(22.94)
2025 release, head-to-head	83	75.9	83	24.1	51.8^{***}	(5.52)

A Variable Definitions

Table A1. Data Dictionary: Outcome Variables

Variable	Definition
`Win`	100 when the user votes for side A, 0 when the user votes for side B. Ties and both-bad votes are excluded, leaving 2,212 battles. This is the outcome in every table unless the table says otherwise.
`WinAll`	100 when the user votes for side A, 0 for every other vote. All 3,060 battles.
`NotLose`	100 unless the user votes for side B, so a tie and a both-bad vote both count as 100. All 3,060 battles.
`Score`	100 for a side A win, 50 for a tie or a both-bad vote, 0 for a side B win.
`Tie, BothBad`	100 when the vote takes that value, 0 otherwise. Used in Table 5 and in Panel C of Table C7.

Table A2. Data Dictionary: Answer Attributes

Variable	Definition
`Confidence`	How assertive and definitive the answer is, 1 to 5. Mean 3.70, two-pass reliability 0.81.
`Actionability`	How concrete and implementable the guidance is, 1 to 5. Mean 3.02, reliability 0.87.
`Accuracy`	Composite of factual correctness and suitability, 1 to 5. Equal to the lower of the two subscores adjusted by at most one point, and capped at 2 when the answer fabricates a fact. Mean 3.63, reliability 0.74.
`FactualCorrectness`	Grades the verifiable claims in the answer, 1 to 5, with 3 when the answer makes no checkable claim. Mean 3.69, reliability 0.75.
`Suitability`	Grades the guidance against the CFP Board planning process for advice questions and against completeness and calibration for information questions, 1 to 5. Mean 3.99, reliability 0.69.
`Refusal`	1 when the answer declines to engage with the question. Mean 0.05.
`RecommendsProfessional`	1 when the answer directs the asker to a financial adviser, planner, accountant or attorney. Mean 0.17.
`Personalized`	1 when the answer is tailored to details the asker supplied. Mean 0.25.
`LogAnswerLength`	Natural log of the answer's length in characters, measured on the full untruncated first answer. 11.0 percent of answers exceed the grader's character limit and 19.0 percent of battles have at least one such answer.

Table A3. Standardization Divisors s_x, the Standard Deviation of Each Within-Battle Difference over All 3,060 Core Battles

Differences in the three graded attributes correlate with one another: ΔConfidence with ΔActionability at 0.41, and ΔConfidence with ΔAccuracy at −0.14. The gap in log answer length correlates 0.34 with the actionability gap.

Difference	s_x
ΔConfidence	1.030
ΔActionability	1.131
ΔAccuracy	1.200
ΔFactual correctness	1.203
ΔSuitability	1.088
ΔLog answer length	0.869

Table A4. Data Dictionary: Fixed Effects, Controls and Sample Indicators

Variable	Definition
`ModelIndicator`	Signed indicator for each of the 106 models: +1 on side A, −1 on side B, 0 when absent. The reference is the model appearing most often in the estimation sample.
`WeekIndicator`	ISO calendar week of the battle, 25 weeks in the decisive sample. The first week is omitted.
`OutOfDomainStrength`	Bradley–Terry log strength of the model, estimated on the non-finance battles of the same release. Enters as a within-battle difference. A single-regressor substitute for the model indicators.
`ModelPair`	The unordered pair of models in the battle, 1,162 pairs, used as a clustering variable.
`Voter, Session`	The anonymized voter identifier, available for the 2024 release only with 520 voters, and the evaluation session, the finest grouping the 2025 release provides with 1,529 sessions.
`English`	1 when the Arena language tag is English. 74.6 percent of core battles.
`SingleTurn`	1 when the conversation has one user turn. 18.5 percent of core battles have more than one, and only the first answer is ever scored.
`SelfIdentifying`	1 when either answer names its own model or maker. 6 battles.
`Verifiability`	How far a diligent reader could check an answer against public references without knowing anything about the asker, assigned by a classifier reading the question text alone: high (1,376 battles), medium (617) or low (1,067). High covers definitions, rules, arithmetic and historical fact; low covers questions turning on the asker's circumstances, preferences or the future.
`AsksForPrediction`	1 when the question asks for a forecast of prices, returns, rates or economic outcomes, which cannot be verified even in principle. 386 battles, none of them high-verifiability.

Table A5. Data Dictionary: Search Arena Variables

Variable	Definition
`Citations`	Number of sources returned with the answer. The difference has a standard deviation of 7.16 across the 427 finance battles.
`InlineMarkers`	Number of distinct visible citation markers such as [1] in the answer text. Standard deviation of the difference 5.09.
`Judge`	The anonymized voter identifier, 230 voters among the decisive votes. Standard errors cluster on it.
`SelfPlay`	27 battles have the same model label on both sides. Their model indicators are all zero, so they identify only the difference and week coefficients. They are kept.

B Examples of Inaccurate Answers That Won

The four battles below are drawn from the two corners in which the winning answer is at least two points more confident or more actionable and at least two points less accurate than its rival, restricted to English single-turn questions short enough to quote. Scores are the grader's confidence, actionability, and accuracy, in that order.

A Bitcoin transaction history. The asker pastes a block-explorer link to a Bitcoin address and asks where the balance went. Side A came from llama-4-maverick-03-26-experimental and scored 5, 5, and 1. It narrates an investigation it could not have performed, reporting that it clicked through to the transaction list and returning a transaction hash, a date of 20 February 2023, a fee of 0.00002734 BTC, a recipient address, and a closing balance. The hash it gives shares its first 42 characters with the hash of the first Bitcoin payment from one user to another, made in January 2009. The two hashes diverge after that point. Side B came from gpt-4.1-mini-2025-04-14 and scored 1, 2, and 4. It says it cannot browse the web and explains how to read the address history on the explorer. **Side A wins.** The grader gave side A a 1 for factual correctness, the rubric's score when fabrication is pervasive.

A house affordability calculation. This is the battle of Figure 3, described in Section 2.3. Side A came from claude-3-5-sonnet-20240620 and scored 4, 5, and 2. Its payment arithmetic is correct, but it then calls a housing cost of 49 percent of gross monthly income within the 36 percent debt-to-income limit it has just quoted. Side B came from gemini-1.5-flash-api-0514 and scored 2, 3, and 5. It declines to give a single number and lists what a lender would look at. **Side A wins.**

A 25-year return projection. The asker holds 27 shares of a semiconductor fund and wants a 25-year projection, "very short and concise, in a table." Side B came from llama-4-maverick-03-26-experimental and scored 4, 5, and 1. It gives four scenarios in a clean table, puts the position at $6,210 today, and then reports that 10 percent a year for 25 years turns it into $13,500. Compounding $6,210 at 10 percent for 25 years gives $67,300, about five times as much. Every row falls short of what its own stated growth rate implies, by a factor that narrows from about five in the first row to

about two in the last. The multiples column is about twice what the dollar column beside it implies, and the plain-English summary underneath contradicts the multiples column, describing the row labeled 4.5 times as more than doubling your money. Side A came from deepseek-v3-0324, scored 3, 3, and 3, and has errors of its own. It uses a share price about four times the one side B uses, and its dollar ranges fall short of what its stated growth rates imply by a factor of about two to four. **Side B wins.**

A ranking of money market funds. The asker wants the top five short-duration USD money market ETFs domiciled in Ireland or Luxembourg, accumulating and liquid, ranked on trailing twelve-month return, with the findings in "an easy to read table." Side A came from mistral-medium-2505 and scored 4, 5, and 2. It returns a table of five funds with tickers, domiciles, trailing returns to a tenth of a percent, expense ratios, assets under management, and a liquidity rating, followed by a ranking of best performer, most liquid, and lowest cost. Trailing returns at that precision require live data that the model did not have. One row lists a "Lyxor USD Overnight Return UCITS ETF" under the ticker USDO. The closest real fund is the Smart Overnight Return UCITS ETF, which trades as SMTC and LXORF. That fund had been rebranded from Lyxor to Amundi in March 2025, before the battle took place. Side B came from claude-sonnet-4-20250514 and scored 1, 3, and 4. It says it cannot give current performance figures because of its knowledge cutoff, then lists the screening criteria, the providers to look at, and four places to find the numbers. **Side A wins.**

In three of the four battles, the losing answer declined to produce what the asker requested, whether an address, a number, or a ranked list.

C Supplementary Estimates

Table C1. Probability That Side A Wins, by the Raw Within-Battle Difference

The numbers behind Figure 2. Each cell of Panel A is the share of decisive votes won by side A among battles with the column's raw difference in 1 to 5 units, and Panel B counts those votes; gaps of three points or more are pooled into the end columns. The sample is the 3,060 core Chatbot Arena finance battles, meaning the advice-seeking and information-seeking questions drawn from the two public preference releases, June to August 2024 and April to July 2025. A battle is one question answered by two anonymous models, with the user voting for one side, calling it a tie, or calling both bad. A decisive vote is a vote for side A or side B; 2,212 of the 3,060 votes are decisive.

Panel A: Side A win rate

	Raw difference, side A minus side B						
	≤ −3	−2	−1	0	+1	+2	≥ +3
ΔConfidence	0.150	0.212	0.405	0.480	0.587	0.677	0.756
ΔActionability	0.226	0.295	0.372	0.479	0.608	0.687	0.750
ΔAccuracy	0.319	0.484	0.487	0.477	0.514	0.524	0.463

Panel B: Decisive votes

	Raw difference, side A minus side B						
	≤ −3	−2	−1	0	+1	+2	≥ +3
ΔConfidence	40	80	393	1,155	404	99	41
ΔActionability	62	132	395	1,039	398	134	52
ΔAccuracy	72	124	439	958	418	147	54

Table C2. The Confidence Gradient at Every Level of Relative Accuracy

The regression counterpart of Figure 6. The decisive votes are partitioned into three bins by the accuracy gap and the vote regression is estimated in each. If confidence only stood in for accuracy, its coefficient would shrink once the sample is held to battles with a given accuracy gap. The outer bins hold about 200 battles each, too few to carry the model indicators, so every column of Panel A uses the same parsimonious control: week fixed effects and each model's Bradley–Terry strength estimated on the non-finance battles of its own release, one regressor in place of a hundred. The accuracy coefficient within a bin is identified only off the truncated variation the bin leaves, and is reported for completeness rather than for interpretation. Panel B reports the same comparison without any controls: the share of decisive votes won by side A at a two-point deficit, at a tie and at a two-point advantage, and the difference between the two ends. The dependent variable takes the value 100 when the user votes for side A and 0 when the user votes for side B, so each coefficient is the change in the probability of a side A win, in percentage points, for a one-standard-deviation difference in the regressor. Parentheses report *t*-statistics based on heteroskedasticity-robust (HC1) standard errors. *, ** and *** indicate statistical significance at the 10%, 5% and 1% level, respectively.

Panel A: Estimates by accuracy bin

		Accuracy bin		
	(1)	(2)	(3)	(4)
Dependent variable: Pr(side A wins), pp	All decisive votes	Side A much less accurate	Similar accuracy	Side A much more accurate
ΔConfidence	4.879***	4.732	4.369***	6.898***
	(4.59)	(1.43)	(3.47)	(2.72)
ΔActionability	4.477***	3.085	4.294***	7.280**
	(4.33)	(0.80)	(3.61)	(2.55)
ΔAccuracy	0.554	14.403**	−1.122	−2.522
	(0.57)	(2.60)	(−0.59)	(−0.33)
ΔOut-of-domain strength	15.355***	13.958***	15.890***	16.105***
	(15.55)	(3.97)	(14.44)	(5.40)
Observations	2,212	196	1,815	201
Fixed effects	Week	Week	Week	Week
Adjusted R^2	0.157	0.214	0.144	0.314

Panel B: Raw win rates behind the gradient

	N	At ≤ −2	At 0	At ≥ +2	Gradient
Δ*Confidence:*					
All decisive votes	2,212	0.192	0.480	0.700	0.508
Side A much less accurate	196	0.067	0.395	0.657	0.590
Similar accuracy	1,815	0.215	0.481	0.704	0.489
Side A much more accurate	201	0.192	0.542	0.857	0.665
Δ*Actionability:*					
All decisive votes	2,212	0.273	0.479	0.704	0.431
Side A much less accurate	196	0.150	0.423	0.667	0.517
Similar accuracy	1,815	0.310	0.475	0.700	0.390
Side A much more accurate	201	0.188	0.567	0.812	0.625

Table C3. Fixed Effects Robustness

Column (1) includes no fixed effects, column (2) adds week fixed effects, column (3) replaces them with model fixed effects, and column (4) includes both, as do columns (5) and (6). The last column replaces the model fixed effects with a single control, each model's Bradley–Terry strength estimated on the non-finance battles of the same release. The dependent variable takes the value 100 when the user votes for side A and 0 when the user votes for side B, so each coefficient is the change in the probability of a side A win, in percentage points, for a one-standard-deviation difference in the regressor. Column (6) enters the raw differences on their 1 to 5 scale, so its coefficients are per one-point difference. Parentheses report *t*-statistics based on heteroskedasticity-robust (HC1) standard errors, except in column (5), where they are clustered by model pair. *, ** and *** indicate statistical significance at the 10%, 5% and 1% level, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dependent variable: Pr(side A wins), pp						Raw 1–5 differences	
ΔConfidence	7.183***	7.159***	4.865***	4.768***	4.768***	4.629***	4.879***
	(6.72)	(6.66)	(4.36)	(4.26)	(4.12)	(4.26)	(4.59)
ΔActionability	8.201***	8.271***	3.869***	3.855***	3.855***	3.408***	4.477***
	(7.80)	(7.86)	(3.48)	(3.47)	(3.42)	(3.47)	(4.33)
ΔAccuracy	3.913***	4.056***	1.043	1.037	1.037	0.865	0.554
	(3.99)	(4.13)	(1.00)	(0.99)	(0.98)	(0.99)	(0.57)
ΔOut-of-domain strength							15.355***
							(15.55)
Constant	48.382***		48.602***				
	(47.21)		(48.98)				
Observations	2,212	2,212	2,212	2,212	2,212	2,212	2,212
Fixed effects	None	Week	Model	Model and week	Model and week	Model and week	Week
Adjusted R^2	0.071	0.074	0.167	0.171	0.171	0.171	0.157

Table C4. Pair Averages Alongside Differences

The difference asks which of the two answers scored higher. The level, the average of the two answers in the battle, asks how good the pair was. Because the pair (difference, level) is a relabelling of the two answers' scores, the two together use everything the grader produced. Levels are centered and scaled by their own standard deviation, so the difference coefficients in column (3) are read at a battle of average quality. The dependent variable takes the value 100 when the user votes for side A and 0 when the user votes for side B, so each coefficient on a difference or a level is the change in the probability of a side A win, in percentage points, for a one-standard-deviation change in the regressor. Each interaction coefficient is the change in the difference coefficient for a one-standard-deviation higher level of the same attribute. Every regressor other than the levels and their interactions is a within-battle difference, side A minus side B, so anything the two answers share drops out. Each difference is divided by its standard deviation across all 3,060 core battles, not within the estimation sample, so coefficients are comparable across columns. Model fixed effects are signed model indicators, +1 when the model answers on side A and −1 when it answers on side B, the Bradley–Terry parameterization, with the model appearing most often in the estimation sample as the reference. Week fixed effects are calendar-week indicators. Parentheses report *t*-statistics based on heteroskedasticity-robust (HC1) standard errors. *, ** and *** indicate statistical significance at the 10%, 5% and 1% level, respectively.

	All questions			Question type	
	(1)	(2)	(3)	(4)	(5)
Dependent variable: Pr(side A wins), pp	Differences	Plus levels	Plus interactions	Advice-seeking	Information-seeking
ΔConfidence	4.768***	4.755***	2.971**	4.063**	4.405***
	(4.26)	(4.25)	(2.23)	(2.25)	(2.82)
ΔActionability	3.855***	3.892***	3.444***	4.198**	3.677**
	(3.47)	(3.49)	(3.05)	(2.41)	(2.30)
ΔAccuracy	1.037	1.040	1.284	−0.733	2.180
	(0.99)	(0.99)	(1.09)	(−0.44)	(1.48)
Level of confidence		0.214	0.336	0.227	−0.186
		(0.20)	(0.32)	(0.13)	(−0.12)
Level of actionability		−0.135	−0.145	3.227	−1.670
		(−0.12)	(−0.13)	(1.34)	(−1.07)
Level of accuracy		−0.648	−0.685	1.122	−1.513
		(−0.61)	(−0.64)	(0.67)	(−1.02)
ΔConfidence × level of confidence			−2.885***		
			(−2.61)		
ΔActionability × level of actionability			4.391***		
			(2.68)		
ΔAccuracy × level of accuracy			0.610		
			(0.42)		
Observations	2,212	2,212	2,212	980	1,232
Fixed effects	Model and week	Model and week	Model and week	Model and week	Model and week
Adjusted R^2	0.171	0.170	0.173	0.189	0.147

Table C5. Alternative Definitions of the Outcome

Each row is a separate regression with model and week fixed effects and a different dependent variable. Using every vote rather than only the decisive ones raises the Chatbot Arena sample from 2,212 to 3,060 battles and the Search Arena sample from 297 to 427. Panel B adds the citation gap and clusters by voter. Each dependent variable takes the value 100 when side A wins, or does not lose, as the row states, and 0 otherwise, with 50 for a tie or a both-bad vote in the score row, so each coefficient is the change in that outcome, in percentage points, for a one-standard-deviation difference in the regressor. Parentheses report *t*-statistics based on heteroskedasticity-robust (HC1) standard errors in Panel A and on standard errors clustered by voter in Panel B. *, ** and *** indicate statistical significance at the 10%, 5% and 1% level, respectively.

Panel A: Chatbot Arena

Dependent variable	N	ΔConfidence	ΔActionability	ΔAccuracy
Side A wins, decisive votes	2,212	4.768***	3.855***	1.037
		(4.26)	(3.47)	(0.99)
Side A does not lose,	3,060	1.966**	3.444***	0.603
all votes		(1.99)	(3.49)	(0.65)
Side A wins, ties and	3,060	4.799***	3.245***	0.823
both-bad count as zero		(5.20)	(3.38)	(0.93)
Side A's score (100, 50, 0),	3,060	3.383***	3.345***	0.713
all votes		(4.08)	(3.86)	(0.91)
Side A does not lose,	2,741	3.301***	3.663***	0.496
both-bad votes dropped		(3.16)	(3.50)	(0.50)

Panel B: Search Arena

Dependent variable	N	ΔCitations	ΔConfidence	ΔActionability	ΔAccuracy
Side A wins, decisive votes	297	4.509	1.981	2.789	7.246**
		(1.11)	(0.61)	(1.07)	(2.56)
Side A does not lose,	427	4.764	0.437	3.446	5.380**
all votes		(1.25)	(0.18)	(1.29)	(2.37)
Side A wins, ties and	427	2.572	4.130	1.591	4.724**
both-bad count as zero		(0.79)	(1.64)	(0.75)	(2.14)
Side A's score (100, 50, 0),	427	3.668	2.284	2.518	5.052***
all votes		(1.16)	(1.02)	(1.16)	(2.72)
Side A does not lose,	382	4.190	0.811	3.467	5.677**
both-bad votes dropped		(1.06)	(0.29)	(1.32)	(2.38)

Table C6. The Linear Model against the Logit

Each panel reports the full specification of its sample, in percentage points per one-standard-deviation difference, first by ordinary least squares and then as the average marginal effects of the logit on the same design, with delta-method standard errors. Panel A is the 2,212 decisive Chatbot Arena votes and Panel B the 297 decisive Search Arena votes with the citation gap included. Parentheses report *t*-statistics based on heteroskedasticity-robust standard errors in Panel A and standard errors clustered by voter in Panel B. *, ** and *** indicate statistical significance at the 10%, 5% and 1% level, respectively.

Panel A: Chatbot Arena

	Linear probability model	Logit, average marginal effect
ΔConfidence	4.768***	5.029***
	(4.26)	(4.49)
ΔActionability	3.855***	4.185***
	(3.47)	(3.73)
ΔAccuracy	1.037	1.091
	(0.99)	(1.06)

Panel B: Search Arena

	Linear probability model	Logit, average marginal effect
ΔCitations	4.509	5.386
	(1.11)	(1.21)
ΔConfidence	1.981	1.869
	(0.61)	(0.61)
ΔActionability	2.789	3.083
	(1.07)	(1.12)
ΔAccuracy	7.246**	7.920***
	(2.56)	(2.67)

Table C7. The Effect of Asker's Question

The prompt is identical for both answers in a battle, so a prompt feature is a battle-level constant and cannot enter as a within-battle difference. Panel A describes the prompts. The detail index used in Panels B and C averages the standardized logs of three counts: first-person words, money references and digits. A prompt is coded as asking for a format when it names a table, a list, bullets, or asks for a short, brief, concise or summarized answer. Panel B interacts each prompt feature with the three attribute differences, one regression per feature, under the same fixed effects as Table 2. Panel C regresses each vote outcome on the three prompt features together with the attribute differences, on all 3,060 votes. In Panel B the dependent variable takes the value 100 when the user votes for side A and 0 when the user votes for side B, and in Panel C each outcome takes the value 100 when the vote takes the column's value and 0 otherwise, so each coefficient is the change in the probability of that outcome, in percentage points, for a one-standard-deviation difference in the regressor. The prompt features are standardised over all 3,060 core battles. Parentheses report *t*-statistics based on heteroskedasticity-robust (HC1) standard errors. *, ** and *** indicate statistical significance at the 10%, 5% and 1% level, respectively.

Panel A: What people ask an AI adviser

	All core battles	Advice-seeking	Information-seeking
Words in the prompt (median)	18	22	15
Words in the prompt (mean)	48.426	46.323	50.028
First-person words per prompt	1.350	1.940	0.901
Money references per prompt	0.400	0.651	0.209
Digits per prompt	8.200	9.829	6.959
Percent asking for a format	6.1	6.3	6.0
Percent with any money reference	8.7	12.4	5.9

Panel B: Interactions with the attribute differences

	Prompt feature		
Dependent variable: Pr(side A wins), pp	Log prompt words	Detail index	Asks for a format
ΔConfidence × prompt feature	0.632	0.369	1.218
	(0.56)	(0.29)	(1.20)
ΔActionability × prompt feature	−0.991	0.758	0.041
	(−1.01)	(0.73)	(0.03)
ΔAccuracy × prompt feature	−0.804	1.164	1.470
	(−0.82)	(1.38)	(1.42)
Observations	2,212	2,212	2,212
Fixed effects	Model and week	Model and week	Model and week

Panel C: Prompt features and each vote outcome

Dependent variable	Side A wins	Tie	Both bad
Log prompt words	0.519	0.836	−0.588
	(0.51)	(0.98)	(−0.83)
Detail index	−0.035	−1.470*	0.289
	(−0.04)	(−1.81)	(0.47)
Asks for a format	−1.289	0.346	0.226
	(−1.51)	(0.46)	(0.36)
Observations	3,060	3,060	3,060
Fixed effects	Model and week	Model and week	Model and week

Table C8. The Two Versions of Llama 4 Maverick

Panel A compares the questions that the two versions of Table 7 drew in the core sample and the rival answers they faced, with a two-sample t-test of each difference. Panel B uses every answer of the two versions in the 2025 release, on all topics. Its measures are computed from the text of the answer and need no grader. The last three measures use English word lists and are computed on English-tagged battles. Column (1) regresses the measure on an indicator for the tuned version with week effects and indicators for English, a follow-up battle, and more than one user message, with standard errors clustered by session. Column (2) is the mean within-battle difference in the 124 battles in which the platform drew the two versions against each other, with a paired t-test. Parentheses report t-statistics. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

Panel A: Questions and rival answers in the core sample

Question or rival answer	Tuned version	Public release	Difference	t-statistic
Advice-seeking question, percent	48.51	48.81	−0.29	(−0.04)
High verifiability, percent	39.60	38.10	1.51	(0.21)
Low verifiability, percent	38.61	44.05	−5.43	(−0.74)
English-tagged, percent	73.27	67.86	5.41	(0.80)
Follow-up battle, percent	13.86	19.05	−5.19	(−0.94)
More than one user message, percent	19.80	17.86	1.94	(0.34)
Log prompt words	3.01	2.70	0.31	(1.92)
Rival's confidence	3.71	3.46	0.25	(1.67)
Rival's actionability	3.29	3.08	0.20	(0.96)
Rival's accuracy	3.57	3.71	−0.14	(−0.92)

Panel B: Text of the answers, 2025 release, all topics

	Mean		Tuned version minus public release	
Text measure of the answer	Tuned version	Public release	(1) All battles	(2) Same question
Log answer length, characters	8.27	7.26	1.02***	0.93***
			(51.74)	(12.03)
Words	543.50	281.97	270.16***	273.27***
			(45.05)	(10.89)
List items	29.09	8.95	20.04***	21.32***
			(49.50)	(10.47)
Numbers	50.48	17.36	33.97***	37.70***
			(32.93)	(10.85)
Any emoji, percent of answers	35.65	3.04	33.41***	26.61***
			(47.97)	(6.42)
Has a table, percent of answers	12.23	2.01	10.24***	13.71***
			(20.20)	(4.42)
Uncertainty words per 1,000 words	6.97	9.55	−2.61***	−2.29*
			(−8.73)	(−1.85)
Weak modal words per 1,000 words	2.66	5.21	−2.60***	−2.92***
			(−14.82)	(−3.77)
Hedge phrases per 1,000 words	0.10	1.22	−1.04***	−2.16
			(−4.58)	(−1.05)
Answers	5,806	5,332	11,138	124 pairs
Answers in English-tagged battles	3,074	2,846	5,920	81 pairs

Table C9. The Tuned Llama 4 Maverick in LMArena's Release

On 8 April 2025, LMArena published 2,382 anonymous battles of the tuned version (llama-4-maverick-03-26-experimental) against 29 other models on all topics, from 28 March to 1 April 2025. Our keyword filter and classifier find 21 finance battles among them, 5 advice-seeking and 16 information-seeking. Panels A and B use a second grader, Claude Opus 5.5, with our two rubrics. In Panel A, column (1) uses LMArena's release, and columns (2) and (3) use the core-sample battles of the 2025 release in which the named version meets another model. Panel B repeats column (1) of Table 7, Panel A, under both graders, with the same 101 and 84 answers. Panel C uses no grader. We estimate the vote's reward for differences in answer length, headers, list items, and bold phrases on the 90,218 decisive votes of the 2025 release that involve neither version, with model and week fixed effects. The predicted style premium applies these weights to each sample's style differences. Panel D uses every battle of LMArena's release and the tuned version's 5,682 battles against other models in the 2025 release. Its word-list measures, as in Appendix Table C8, use English-tagged battles, and every difference between the tuned version and its rival in it is significant at the 1% level. Parentheses report *t*-statistics, paired within battle in Panel A and heteroskedasticity-robust (HC1) in Panel B. *, ** and *** indicate statistical significance at the 10%, 5% and 1% level, respectively.

Panel A: Within-battle gaps against rivals, second grader

Score, own answer minus rival's	(1) LMArena release Tuned version	(2) 2025 release Tuned version	(3) 2025 release Public release
Confidence	1.05^{***}	0.57^{***}	-0.14
	(4.93)	(7.08)	(-1.37)
Actionability	0.48^{**}	0.55^{***}	-0.27^{***}
	(2.68)	(5.78)	(-3.09)
Accuracy	-1.33^{***}	-0.85^{***}	-0.17
	(-7.68)	(-7.45)	(-1.37)
Battles	21	98	81

Panel B: The 2025 comparison of Table 7 under two graders

Tuned version minus public release	Paper's grader	Second grader
Confidence	0.706^{***}	0.934^{***}
	(5.82)	(8.11)
Actionability	1.145^{***}	1.215^{***}
	(5.55)	(7.24)
Accuracy	-0.555^{***}	-0.646^{***}
	(-3.19)	(-4.74)

Panel C: Votes and the predicted premium for style

Sample, all topics	Decisive votes	Win rate, percent	Predicted style premium, pp
LMArena release, tuned version	1,774	66.5	15.5
2025 release, tuned version	4,230	61.6	11.3
2025 release, public release	3,817	36.8	-3.0
2025 release, tuned vs public	83	75.9	14.0

Panel D: The text of the answers

Mean per answer, all topics	LMArena release Tuned version	LMArena release Rival	2025 release Tuned version	2025 release Rival
Words	685	286	542	334
Any emoji, percent of answers	74.4	12.8	35.7	22.7
Uncertainty words per 1,000 words	6.67	10.70	6.93	9.26
Hedge phrases per 1,000 words	0.06	0.75	0.10	0.68