Common Models, Different Beliefs

AI and Financial Disagreement

Jonas Theill Bøjstrup

jobo@econ.au.dk

Abstract

Presentation produces persistent differences in the financial forecasts supplied by a common large language model. A controlled experiment varies numerical, graphical and textual representations of simulated market histories. Forecast differences recur across independent requests, and text summaries produce the strongest extrapolation from earlier returns. Reviewing another representation narrows disagreement, but forecasts still imply different investment positions in about 23% of comparisons after both representations are available. More intensive reasoning improves the expected payoff of implied positions most when earlier and latest returns conflict, even as average squared probability error rises.

Keywords: Artificial intelligence; financial beliefs; investor disagreement; information representation.

JEL Classification: D83, D84, G11, G17, G41.

Department of Economics and Business Economics
Aarhus University
Center for Research in Energy: Economics and Markets (CoRE)

1 Introduction

Delegating financial analysis to artificial intelligence (AI) leaves an important choice with the investor: how to present the information. Large language models (LLMs) can interpret numerical data, images and text, allowing investors to use a common model while approaching financial information in different ways. As AI becomes part of investment research (Cao et al., 2024), these choices can become a source of differences in the assessments supplied to investors. This paper shows that representation produces recurring differences in financial forecasts and implied investment positions within a common model. Those differences persist after the forecasts are revised using the same pair of representations.

An investor faced with conflicting assessments can return to the underlying evidence to evaluate the basis for each view. When expectations depend on how a history is presented (Glaser et al., 2019; Chen et al., 2025), examining another representation offers a way to recognize features overlooked in the initial assessment. If access to those features reconciles the forecasts, the initial disagreement can be resolved by reviewing the available material. If the starting view continues to influence the revised assessment, presentation has consequences that survive that review. Investors can then obtain different interpretations even after consulting the same material through the same model. Establishing how far common access reconciles these views helps explain why financial disagreement persists and what sharing the underlying evidence accomplishes.

Seeking a separately formed opinion offers another way to evaluate an initial assessment. It provides an alternative interpretation to compare with the original and can improve a combined forecast when the errors in the two assessments offset one another. The value of either reviewing the data or obtaining another opinion depends on how the resulting beliefs change investment positions. Forecasts can move closer together while continuing to imply different positions, and a revision that reduces average forecast error need not improve the decision to buy, sell or stay out. Conversely, a change in how the history is

interpreted can improve those decisions even when average forecast error rises. Following disagreement through revision therefore connects the formation of financial beliefs to the value of the decisions they support.

To examine how these differences arise and persist, a controlled experiment follows forecasts for 1,000 simulated price histories spanning different market states. A single LLM receives twenty days of prices and volume as a numerical table, a chart image or a text summary; the image and text are derived from the table. It reports the probability of a positive return, expected return and the standard deviation of returns at one-, five- and twenty-day horizons. The price-generating process is never disclosed to the model. Each initial request is repeated independently to distinguish recurring representation differences from variation in responses to the same input. Subsequent requests supply another representation and ask the model to revise its forecast. The model is run with its reasoning setting enabled and disabled, varying the computation allocated before an answer is produced. A separate revision supplies a forecast generated independently from the same or another representation.

The initial forecasts disagree substantially about individual histories despite similar average probabilities across representations. With reasoning enabled, the root mean squared probability gap is 8.3 percentage points across representations, compared with 5.7 points between repeated forecasts using the same representation. The differences recur in independent requests under both reasoning settings. Their financial consequences are measured by applying a common decision rule to a claim that pays one if the future return is positive. Each probability implies a long, flat or short position at a common price and trading cost. Positions differ in 25.5% of representation comparisons with reasoning and 31.5% without it. Averaging forecasts across histories therefore conceals substantial disagreement about the position to take in a given history.

Forecasts respond differently to the price movements within each history. Separating the return before the final day from that day's move shows that text summaries produce the strongest positive response to the earlier return, in both the probability of a gain and the

expected return. Access to another representation narrows these response differences and brings forecasts closer together, but the starting view remains evident. For example, a chart-based forecast is revised after the text summary is supplied, while a summary-based forecast is revised after the chart is supplied. Both requests contain the same market representations, alongside their respective initial forecasts. Across representation pairs, the probability gap falls from 8.3 to 6.3 percentage points with reasoning and from 7.0 to 4.7 points without it. The remaining differences align with the initial differences measured in independent requests. Expected-return and risk forecasts also converge, while implied positions still differ in about 23% of comparisons under either reasoning setting. The influence of the initial presentation therefore survives access to both views of the history.

More intensive reasoning changes the response to past returns and improves the implied investment decisions. Enabling reasoning increases the average response to the earlier return and reduces the response to the final day's return. Expected payoff rises by 67 basis points of claim face value. The gain is 114 basis points larger when the two returns have opposite signs than when they have the same sign. Reasoning improves these decisions even as it increases average squared probability error. Simple averaging accounts for the probability-accuracy gain from a separately generated forecast based on another representation; with reasoning enabled, the model's further revision gives back part of that gain. The accuracy advantage over averaging forecasts from the same representation becomes much smaller after adjusting for systematic probability errors, and the revised forecasts yield no statistically discernible increase in expected payoff.

The findings identify presentation as a source of persistent differences in the financial assessments supplied by AI. A common model and access to the same market representations leave room for different views of returns and risk, with consequences for investment positions. For investors and providers of financial analysis, the presentation of data shapes the investment views supplied by a model, including those formed after the data are reviewed in another form. The results also distinguish the benefits of these changes in analysis: broader access brings beliefs closer together, averaging improves probability scores, and more intensive reasoning improves the expected payoff of the

implied positions. Agreement, forecast accuracy and financial value capture different consequences of how information is processed.

1.1 Literature review

The paper contributes to research on the origins of financial disagreement. Different interpretations of public information can sustain differences in beliefs and trading (Harris and Raviv, 1993; Kandel and Pearson, 1995; Banerjee and Kremer, 2010; Banerjee, 2011). Cookson and Niessner (2020) distinguish disagreement within and across investors' stated investment approaches, while Mondria et al. (2022) separate the cost of interpreting asset prices from observing them. The experiment identifies recurring differences within a common forecasting model and measures how much of the initial disagreement remains after both representations are available. It connects the formation of heterogeneous assessments to their persistence under common access. This evidence complements models of disclosure, information acquisition and specialization (Diamond, 1985; Van Nieuwerburgh and Veldkamp, 2009; Goldstein and Yang, 2015) by documenting a continuing dependence on the starting representation after access to the market inputs has been equalized.[1]

The closest evidence on LLM financial beliefs comes from Chen et al. (2025), who document extrapolation, miscalibration and sensitivity to numerical and graphical presentations of return histories, and Bini et al. (2026), who examine biases in beliefs and preferences across model families, including forecasts of autoregressive processes and investment choices. The contribution here is to establish which representation differences recur, follow the same assessments through revision, and measure their consequences for positions. Reciprocal exposure to both representations tests whether common access reconciles those assessments; a separately generated forecast tests the value of combining opinions. This focus on belief formation complements research extracting investment signals from numerical predictors, price images, news and financial statements (Gu et al., 2020; Jiang et al., 2023; Lopez-Lira and Tang, 2026; Kim et al., 2024).[2]

[1] For the aggregation of dispersed information into prices, see Grossman and Stiglitz (1980); Hellwig (1980); Diamond and Verrecchia (1981).

[2] Related applications use LLMs to measure human expectations from historical news (Bybee, 2025)

The evidence also connects presentation to the use of past returns and the resulting financial choices. Processing constraints shape responses to disclosures and the allocation of attention (Hirshleifer and Teoh, 2003; Peng and Xiong, 2006); prominent news placement changes the speed with which public information enters prices (Fedyk, 2024). Glaser et al. (2019) show that price and return presentations change human expectations, while Hartzmark and Shue (2018) document the influence of recent comparisons on responses to earnings news. The present results identify differences in responses to earlier and latest returns across representations and show how those differences change when another representation or more computation is supplied. They connect evidence on extrapolative expectations (Greenwood and Shleifer, 2014; Cassella and Gulen, 2018; Da et al., 2021) to the financial consequences of changing the way a history is processed. Under a common investment rule, reasoning improves decisions most when the two return movements conflict. This complements evidence linking beliefs to portfolios (Giglio et al., 2021) and the presentation of returns to risk taking (Beshears et al., 2017), as well as research on automated advice (D'Acunto et al., 2019) and firms' responses to machine readers (Cao et al., 2023).

Finally, the second-opinion treatment connects disagreement to the use of others' assessments. Social interaction influences participation, portfolio choices and information acquisition (Hong et al., 2004; Pool et al., 2015; Halim et al., 2019). Online discussions convey investment information and can reinforce existing views (Antweiler and Frank, 2004; Chen et al., 2014; Cookson et al., 2023). The experiment compares opinions based on the same and different representations, with each format used equally often for the initial and additional forecast. Using forecast averaging as a benchmark (Bates and Granger, 1969), it separates the benefit of offsetting errors from the model's subsequent revision and the resulting change in positions. The dependence between forecasts enters this accounting directly, connecting the analysis to research on correlation in belief updating (Enke and Zimmermann, 2019). This comparison identifies what a different interpretation contributes before asking whether further model analysis adds financial value.

and disagreement among investor personas (Bhagwat et al., 2026). Here, the object is the forecast supplied by the model itself, without assigning different investor identities.

2 Experimental Design and Measurement

This section specifies the financial environment, the information available in each forecast request, and the comparisons used to measure disagreement and its persistence. It then defines how forecasts are evaluated and translated into investment positions.

2.1 Financial environment

The sample comprises 1,000 independent price histories. Each forecast uses twenty completed daily observations and concerns returns over the next one, five and twenty days. Simulating returns permits comparisons across specified market conditions, as in the controlled forecasting exercises of Chen et al. (2025), and evaluation over repeated future realizations. The generating process, its parameters and the underlying market state are never disclosed to the model.

The environment contains four latent states: continuation, reversal, no predictability and stress. Continuation and reversal allow the relation between past and subsequent returns to change sign (De Bondt and Thaler, 1985; Jegadeesh, 1990; Jegadeesh and Titman, 1993). The no-predictability state applies the benchmark of a constant conditional mean to daily log returns (Fama, 1970). Stress combines elevated volatility with large downward shocks, as observed during market disruptions (Schwert, 1990). State transitions follow a Markov chain, as in the regime-switching framework of Hamilton (1989). The experiment assigns stationary weights of 25%, 25%, 40% and 10% to the four states and samples 250, 250, 400 and 100 forecast origins, respectively.

Let r_t denote the close-to-close log return on date t, μ_0 the common daily drift and s_t the latent market state. Write $\widetilde{r}_t = r_t - \mu_0$ for the return net of this drift. Returns satisfy

$$r_t = \mu_0 + m_t(s_t) + \sqrt{v_t}\,\varepsilon_t + J_t, \tag{2.1}$$

$$m_t(s) = \delta_s + \phi_s \widetilde{r}_{t-1} + \beta_{5,s}\frac{1}{5}\sum_{u=1}^{5}\widetilde{r}_{t-u} + \beta_{20,s}\frac{1}{20}\sum_{u=1}^{20}\widetilde{r}_{t-u}. \tag{2.2}$$

The drift component $m_t(s)$ combines a state-specific offset δ_s with dependence on past returns. The coefficients ϕ_s, $\beta_{5,s}$ and $\beta_{20,s}$ govern the responses to the latest return and the averages of the preceding five and twenty daily returns, respectively. They are positive in continuation, negative in reversal and zero in the other two states.

The variance of the Gaussian return component is v_t, drawn from a state-dependent lognormal distribution motivated by evidence on realized volatility (Andersen et al., 2001). Its standardized shock, $\varepsilon_t \sim N(0,1)$, is independent across dates and of the state process, variance draws and jumps. The realized jump return J_t is zero when no jump occurs and normally distributed otherwise, with arrival probabilities and jump-size distributions varying by state. This component accommodates discontinuous price movements (Merton, 1976). Stress has higher variance, more frequent jumps and a negative mean jump size. The drift offset satisfies $\delta_s = -\mathbb{E}[J_t \mid s_t = s]$, compensating expected jumps. Conditional on the current state and past returns, the mean log return is therefore μ_0 in both the no-predictability and stress states.

The closing price C_t evolves as $C_t = C_{t-1}\exp(r_t)$. Open, high, low and volume are generated from returns, variance and jumps. Parameters were fixed before collecting forecasts. Appendix B reports the transition matrix, parameter choices, observation equations and pre-experiment validation. Taking date zero as the forecast origin, a common future path determines the cumulative log return ℓ_h, simple return R_h and positive-return indicator D_h at horizon h:

$$\ell_h = \sum_{t=1}^{h} r_t, \qquad R_h = e^{\ell_h} - 1, \qquad D_h = \mathbf{1}\{R_h > 0\}, \quad h \in \{1, 5, 20\}. \tag{2.3}$$

2.2 Representations and forecast revision

Each history is presented as a numerical table, a chart image or a text summary. Prices are rebased to an initial displayed close of 100 and rounded to two decimals; volume in millions is rounded to three decimals. The table contains twenty daily observations of open, high, low, close and volume. The chart displays prices as candlesticks and volume

as bars. The text summary reports price changes, return volatility, price ranges and volume statistics in fixed wording and order. Both are constructed from the rounded table. Figure 2.1 shows the three inputs for one history; Appendix C specifies their construction.

The forecast questions remain the same across representations. At each horizon, the model reports the probability of a positive return and the mean and standard deviation of the cumulative simple return. These forecasts describe the anticipated direction, size and risk of returns. Holding the questions fixed separates changes in the displayed information from changes in how expectations are elicited, a distinction central to Glaser et al. (2019).

The forecaster is GPT-5.6 Terra. It is evaluated with medium reasoning effort and with reasoning disabled, varying the computation allocated before the answer is produced. Instructions, forecast questions and required response fields are otherwise unchanged. Each initial request is repeated independently to measure variation between responses to the same input.

To review an initial assessment, the model receives its original representation and forecast together with another representation of the history. It receives no forecast based on the added representation. All six ordered pairs are observed: for example, a chart-based forecast is revised after the text summary is supplied, and a summary-based forecast is revised after the chart is supplied. Comparing these reciprocal revisions measures disagreement after access to the same pair of representations, with each request retaining its original forecast and the representations' starting roles.

Let X denote the complete rounded history and g_j the deterministic transformation producing representation $X_j = g_j(X)$. The table preserves X; the chart and summary retain selected features. The table is therefore at least as informative as either transformation in the sense of Blackwell (1953). Adding a chart or summary to the table supplies no additional information about the history, so its effect on forecasts reflects how the available information is processed. Adding the table to a summary also supplies omitted detail. Under a common prior and likelihood, Bayesian conditioning on representations (X_j, X_k) gives the same probability through either route. Reciprocal

Figure 2.1: Three representations of the same price history

Numerical table

Day	Open	High	Low	Close	Volume
−19	100.80	101.45	99.08	100.00	3.951
−18	101.70	103.85	99.14	100.34	15.360
−17	100.53	101.47	100.08	101.03	8.253
−16	100.82	101.43	100.31	100.61	2.385
−15	100.60	103.37	99.27	102.08	0.870
−14	102.01	102.62	100.56	101.07	0.806
−13	101.64	102.52	95.11	98.43	3.054
−12	98.95	102.83	95.09	100.96	3.275
−11	101.82	104.64	101.21	103.15	4.133
−10	103.42	104.16	102.68	103.35	2.351
−9	103.39	104.29	103.23	103.72	1.454
−8	103.92	105.40	100.97	102.73	1.044
−7	102.98	103.67	102.58	103.35	0.322
−6	103.12	104.20	100.60	101.77	1.444
−5	101.08	103.97	98.01	103.18	3.386
−4	104.61	105.26	101.21	102.71	22.814
−3	103.59	104.60	103.15	104.05	8.051
−2	104.38	105.42	103.40	105.00	2.657
−1	105.16	106.26	103.24	103.99	3.749
0	103.91	106.52	99.59	100.57	5.898

Candlestick and volume chart



Factual market summary

History: 20 daily observations.
Closing-price index: 100.00 to 100.57 (+0.57%).
Most recent close-to-close change: -3.29%.
Positive close-to-close changes: 11 of 19.
Daily close-return volatility: 1.53%.
Largest and smallest daily changes: +2.57% and -3.29%.
Displayed low and high: 95.09 and 106.52; mean daily high-to-low range: 3.50%.
Volume: latest 5.898 million; window average 4.763 million; latest-to-average ratio 1.24.

Note: The three inputs reproduce the matched example retained in the original experimental design. Prices are indices with the first close equal to 100; volume is in millions. Day 0 is the forecast origin. Chart positions 1–20 correspond to table dates −19 through 0. The chart is the original 768-by-512-pixel input image. Table values and summary wording reproduce the retained example; typography is adapted for display.

revision measures whether the starting representation remains reflected in forecasts once both are available.

Consultation allows the forecaster to reconsider its assessment using another interpretation of the history. The model retains its original representation and forecast and receives a labeled peer report from an independent repetition, without the peer's representation. Each of the three formats supplies both initial and peer forecasts, yielding three same-format and six different-format pairs. Receiver and peer frequencies are balanced within each group, so their comparison holds the mix of constituent forecasts fixed. The contribution of consultation can then be evaluated relative to simply averaging the two forecasts. Together with separate copying controls, the design produces 54,000 recorded requests. Appendix D reports fallback counts and the complete-case sample.

2.3 Measurement of beliefs and disagreement

Disagreement is measured by gaps between forecasts for the same history. Root mean squared gaps summarize their magnitude for probabilities, expected returns and stated risk. Comparisons between representations are evaluated alongside differences between independent requests using the same representation.

Independent repetitions also identify the recurring component of probability disagreement. For a given horizon and reasoning condition, let $p_j^{(1)}$ and $p_j^{(2)}$ denote the two initial probabilities from representation j, with common conditional mean $b_j(X) = \mathbb{E}[p_j^{(1)} \mid X] = \mathbb{E}[p_j^{(2)} \mid X]$. Conditional independence of the repetitions implies

$$\mathbb{E}\Big[(p_j^{(1)} - p_k^{(1)})(p_j^{(2)} - p_k^{(2)})\Big] = \mathbb{E}\Big[(b_j(X) - b_k(X))^2\Big]. \tag{2.4}$$

The expected product measures differences in the average forecast for a history, removing independent response variation. Its sample counterpart, averaged across histories, the three representation pairs and horizons, is denoted $\mathcal{S}$.

After another representation is supplied, convergence is measured by the reduction in

reciprocal forecast gaps. Persistence is measured by their alignment with the independent initial differences. Let $d_{jk}^{(+)}$ be the revised forecast from route $j \to k$ minus that from route $k \to j$. Define

$$\mathcal{P} = \left\langle d_{jk}^{(+)} (p_j^{(2)} - p_k^{(2)}) \right\rangle, \tag{2.5}$$

where brackets denote the same average used for $\mathcal{S}$. A positive $\mathcal{P}$ means that final disagreement remains aligned, on average, with the initial representation differences. The change $\mathcal{S} - \mathcal{P}$ measures the reduction in that alignment. Appendix A derives the repeated-gap identity and related loss decompositions; Appendix D gives the estimators and units.

Responses to past returns describe which movements in the history are associated with these beliefs. Following research relating expectations to past returns (Greenwood and Shleifer, 2014; Chen et al., 2025), the analysis regresses forecasts on the earlier return and the latest daily return. The earlier return covers the first eighteen close-to-close intervals; the latest return covers the final interval. Both log returns enter each regression and are standardized over the full sample. Coefficients are estimated separately by representation, reasoning condition and horizon. Histories are classified as conflicting when the two returns have opposite signs and agreeing when their signs coincide; exact-zero cases are excluded from this classification. Applying the specification to probabilities, expected returns and stated risk connects representation differences to the use of the same price movements.

2.4 Financial choices and forecast evaluation

The financial consequences of disagreement are measured by applying a common investment rule to every probability forecast. The asset is a claim paying one if the future return is positive and zero otherwise. Its expected payoff under the reported belief is therefore the reported probability p. At price Q_h, a risk-neutral investor chooses a long, flat or short

position $a(p; Q_h, k) \in \{1, 0, -1\}$, with cost k per unit of position:

$$a(p; Q_h, k) = \begin{cases} 1, & p > Q_h + k, \\ 0, & Q_h - k \leq p \leq Q_h + k, \\ -1, & p < Q_h - k. \end{cases} \tag{2.6}$$

Prices are fixed at the positive-return frequencies from a separate pre-experiment simulation of one million histories: 0.523, 0.539 and 0.565 at the three horizons. The main cost is $k = 0.01$, creating an interval around the price in which no position is taken. Holding the claim, price, cost and available positions fixed makes differences in choices depend on the reported beliefs.

Expected payoffs are evaluated by averaging over possible futures from each forecast origin. Let q_{eh} be the probability of a positive return conditional on episode e's complete origin state under the generating process. The expected net payoff of the chosen position is

$$\pi_{eh}(p) = a(p; Q_h, k)(q_{eh} - Q_h) - k|a(p; Q_h, k)|. \tag{2.7}$$

The evaluator estimates q_{eh} using 8,192 stored future continuations. The model receives neither this probability nor the latent state, and its reported forecast alone determines the position. This evaluation integrates out future return realizations conditional on the origin state; it does not estimate the posterior based only on the displayed history. Payoff differences are expressed in basis points of claim face value. Appendix D reports exact prices and numerical uncertainty.

Probability accuracy is measured by Brier loss, $(D_h - p_h)^2$ (Brier, 1950), a strictly proper scoring rule (Gneiting and Raftery, 2007). Conditional expected loss averages this score over the event using q_{eh}. Squared error evaluates probabilities throughout their range, whereas positions change only when a forecast crosses a decision boundary. The two measures therefore assess different consequences of a forecast revision.

For consultation, equal averaging provides a benchmark for the contribution of combining

forecasts (Bates and Granger, 1969). For a given horizon, write $D = D_h$ and let $\bar{p}_{jk} = (p_j^{(1)} + p_k^{(2)})/2$ be the equal pool. Its loss satisfies

$$(D - \bar{p}_{jk})^2 = \frac{(D - p_j^{(1)})^2 + (D - p_k^{(2)})^2}{2} - \frac{(p_j^{(1)} - p_k^{(2)})^2}{4}. \tag{2.8}$$

Balanced receiver and peer frequencies equate average constituent loss across same-format and different-format pairs. Same-format minus different-format equal-pool loss is exactly $\mathcal{S}/4$ in the sample. Comparing the final forecast with the equal pool measures the contribution of the model's subsequent revision. Appendix A gives the derivation.

Probability calibration assesses how systematic probability errors affect the comparisons. It regresses the realized binary outcome on the reported probability and an intercept. A separate adjustment fits reasoning probabilities to probabilities without reasoning, with an intercept and a nonnegative slope. This fit uses no realized outcomes and asks whether a common change in probability level and scale accounts for the change in payoffs. Both procedures are estimated on training histories and evaluated on held-out histories using five folds; Appendix D provides the specifications.

All comparisons are paired by history. Each horizon is evaluated separately before equal averaging. Inference resamples complete histories within market states, retaining all linked forecasts, and reported intervals are simultaneous within the specified contrast family. The original confirmatory contrasts evaluate revisions following another representation or a peer forecast in normalized return loss, with Brier loss as a secondary outcome. Subsequent analyses of disagreement, positions, calibration and convergence use the same collected forecasts. Appendix E reports the original contrasts and complete-case estimates; Appendix H defines the return loss and its population moment properties.

3 Empirical Analysis

3.1 Initial financial beliefs

Table 3.1 reports the initial probability forecasts, expected returns and stated risk by representation, reasoning condition and horizon. Average probabilities are similar across representations. With reasoning, the mean probability of a positive twenty-day return ranges from 52.0% for chart images to 53.7% for text summaries. Their dispersion across histories differs substantially: the standard deviation is 13.5 percentage points for charts, 16.2 for numerical tables and 21.6 for text summaries.

Text summaries produce the most dispersed probabilities at every horizon under both reasoning conditions. Enabling reasoning also increases dispersion within every representation and horizon. At twenty days, for example, the standard deviation of probabilities from numerical tables rises from 9.3 to 16.2 percentage points.

Expected returns and stated risk also differ across representations. With reasoning, text summaries yield the highest average twenty-day expected return, 0.84%, compared with 0.68% for tables and 0.44% for charts. Yet summaries yield the lowest average stated return standard deviation, 6.22%, compared with 6.68% for tables and 6.46% for charts. The greater dispersion of summary-based probabilities across histories thus coexists with lower average stated uncertainty about future returns.

3.2 Recurring disagreement across representations

Forecasts for the same price history differ substantially across representations. Table 3.2 compares these differences with those between independent requests using the same representation. With reasoning, the root mean squared probability gap is 8.3 percentage points across representations and 5.7 points within a representation. Without reasoning, the corresponding gaps are 7.0 and 2.9 points. In both conditions, changing the representation produces larger differences than repeating the request.

Table 3.1: Descriptive statistics of initial forecasts

Reasoning	Format	Horizon	Mean probability (%)	SD of probability (pp)	Mean expected return (%)	Mean stated return SD (%)	Brier loss
Reasoning	Raw table	1 day	50.88	6.01	0.03	1.42	0.250
		5 days	52.01	10.88	0.17	3.21	0.251
		20 days	53.41	16.19	0.68	6.68	0.258
	Chart	1 day	50.55	5.17	0.03	1.38	0.250
		5 days	51.33	9.18	0.13	3.13	0.248
		20 days	52.04	13.46	0.44	6.46	0.253
	Summary	1 day	51.05	7.76	0.04	1.34	0.253
		5 days	52.29	14.21	0.19	3.02	0.255
		20 days	53.70	21.59	0.84	6.22	0.277
No reasoning	Raw table	1 day	49.96	3.75	0.01	1.89	0.251
		5 days	49.82	5.91	0.10	4.20	0.250
		20 days	49.84	9.32	0.44	8.54	0.248
	Chart	1 day	49.71	3.39	0.00	1.32	0.251
		5 days	50.12	4.98	0.07	2.96	0.249
		20 days	50.66	7.72	0.36	6.04	0.248
	Summary	1 day	50.52	6.95	0.04	1.34	0.252
		5 days	50.84	10.18	0.23	3.01	0.252
		20 days	51.38	15.16	1.03	6.20	0.259

Note: Each row contains 1,000 first forecasts. Probability refers to a positive cumulative return. Probability SD measures dispersion of reported probabilities across histories; stated return SD is the model's forecast of return risk. Expected returns and stated return standard deviations are in cumulative simple-return units. The realized frequencies of positive returns are 52.2%, 57.4% and 57.2% at one, five and twenty days. Brier loss is the mean squared error of the stated probability. All entries use original reports before calibration or revision.

Independent repetitions establish that these differences recur. The repeated-gap product in equation (2.4) is 36.3 squared percentage points with reasoning and 41.1 without reasoning; both simultaneous 95% intervals exclude zero. Representation differences therefore have a recurring direction across requests, beyond independent variation in the model's responses. Both pairs involving text summaries have larger recurring differences than the table–chart pair (Table D.2). These results locate a source of disagreement within a common model, complementing the differences across investment approaches documented by Cookson and Niessner (2020).

The differences in beliefs change the positions implied by the forecasts. Under the common investment rule, representations yield different long, flat or short positions in 25.5% of comparisons with reasoning and 31.5% without it. Similar average probabilities across representations therefore conceal disagreement about the position to take in roughly one-quarter to one-third of comparisons.

3.3 Responses to past returns

Forecasts respond positively to both earlier and latest returns, consistent with the extrapolative expectations documented in investor surveys (Greenwood and Shleifer, 2014) and LLM forecasts (Chen et al., 2025). Table 3.3 shows how these responses differ across representations, using the joint specification in Section 2.3.

Text summaries produce the strongest response to the earlier return at every horizon under both reasoning conditions. With reasoning, a one-standard-deviation increase in the earlier return, holding the latest return fixed, is associated with an increase of 12.4 percentage points in the probability of a gain for summaries, averaged across horizons. The corresponding responses are 8.1 points for tables and 6.8 for charts. Without reasoning, they are 8.1, 3.2 and 2.7 points, respectively. Independent repetitions reproduce this ordering and yield closely similar coefficients (Table F.1).

Reasoning increases the average response to the earlier return and reduces the response to the latest return in all three representations. For numerical tables, the earlier-return

Table 3.2: Representation and recurring disagreement

Reasoning	Horizon	Within-format RMS (pp)	Across-format RMS (pp)	Repeated-gap product (pp^2)	Position disagreement (%)
Reasoning	1 day	3.3	4.5	9.7	32.3
	5 days	5.1	7.4	29.7	23.8
	20 days	7.8	11.4	69.6	20.3
	Average	5.7	8.3	36.3	25.5
	95% interval			[30.8, 41.8]	
No reasoning	1 day	1.8	4.8	20.0	36.2
	5 days	2.5	6.3	33.1	29.6
	20 days	3.9	9.3	70.1	28.7
	Average	2.9	7.0	41.1	31.5
	95% interval			[36.7, 45.5]	

Note: Within-format disagreement compares first and independent-repeat forecasts of the same representation. Across-format disagreement compares first and independent-repeat forecasts of different representations. The repeated-gap product multiplies each representation difference by its counterpart in the independent repeat and averages the three unordered pairs. It equals across-format minus within-format mean squared disagreement. RMS takes the square root after averaging squared gaps. Position disagreement uses first forecasts and records different long, flat or short positions at a 100 bp cost. Brackets report simultaneous 95% intervals for the horizon-average repeated-gap products from the eight-contrast disagreement family. Tables D.1 and D.2 report the full interval family and individual representation pairs.

coefficient rises from 3.2 to 8.1 percentage points per standard deviation, while the latest-return coefficient falls from 3.5 to 2.5 points.[3]

Expected-return forecasts exhibit the same representation ordering. At twenty days, a one-standard-deviation increase in the earlier return is associated with an increase in the expected return of 4.20 percentage points for text summaries, 2.68 for tables and 2.16 for charts when reasoning is enabled. Without reasoning, the coefficients are 4.19, 1.49 and 1.13 points. The stronger extrapolation from text therefore appears in both the probability of a gain and the expected return. Table F.2 also reports the responses of stated risk.

3.4 Belief revision and persistence

Access to another representation brings forecasts closer together while leaving the starting presentation reflected in the revised assessment. Table 3.4 compares reciprocal revisions: forecasts begin with different representations and then receive the other, giving both access to the same pair of market inputs. The root mean squared probability gap falls from 8.3 to 6.3 percentage points with reasoning and from 7.0 to 4.7 points without reasoning.

The remaining differences retain the direction of the initial representation differences. Their alignment with the independent initial forecasts, measured by $\mathcal{P}$, is 18.1 squared percentage points with reasoning and 23.8 without it, compared with initial repeated-gap products of 36.3 and 41.1. Both final products and both reductions are positive throughout their simultaneous 95% intervals (Table D.1).

These differences continue to affect positions. After revision, the two routes imply different long, flat or short positions in 22.7% of comparisons with reasoning and 23.2% without it. The reductions are 2.8 and 8.3 percentage points, respectively.

Expected-return and risk forecasts also converge. With reasoning, the twenty-day expected-return gap falls from 2.73 to 1.78 percentage points, and the stated-risk gap falls from 1.51

[3]Table F.3 reports results and simultaneous intervals for the related specification that replaces the earlier return with the full displayed return, retaining the latest daily return as the second regressor.

Table 3.3: Past returns and financial beliefs

Reasoning / format	1 day: earlier	1 day: latest	5 days: earlier	5 days: latest	20 days: earlier	20 days: latest
Panel A. Probability forecasts (probability pp per SD of past return)						
Reasoning / Raw table	4.32	1.19	8.07	2.53	11.92	3.66
Reasoning / Chart	3.69	0.98	6.84	2.03	9.97	2.94
Reasoning / Summary	6.64	1.57	12.29	2.91	18.14	4.22
No reasoning / Raw table	1.40	2.40	3.05	3.24	5.04	4.74
No reasoning / Chart	0.89	1.78	2.72	2.29	4.61	3.25
No reasoning / Summary	4.99	2.29	7.78	3.06	11.50	4.41
Panel B. Expected-return forecasts (return pp per SD of past return)						
Reasoning / Raw table	0.17	0.04	0.80	0.23	2.68	0.75
Reasoning / Chart	0.15	0.03	0.68	0.20	2.16	0.64
Reasoning / Summary	0.23	0.05	1.15	0.25	4.20	0.85
No reasoning / Raw table	0.07	0.09	0.39	0.39	1.49	1.43
No reasoning / Chart	0.04	0.06	0.29	0.23	1.13	0.80
No reasoning / Summary	0.21	0.08	1.06	0.38	4.19	1.50

Note: Each regression includes an intercept and both standardized past returns, as specified in Section 2.3. Earlier returns cover the first eighteen close-to-close intervals; the latest return covers the final interval. Each regressor is standardized by its own full-sample standard deviation. All rows use first forecasts. Table F.1 reports independent-repeat probability slopes. Table F.2 includes the slopes of stated return standard deviations.

to 1.07 points. Both dimensions also converge without reasoning. Substantial differences remain in pairs involving the numerical table, where both final requests contain the complete displayed price and volume panel (Table G.5).

Responses to earlier returns become more similar, particularly in comparisons involving text summaries. With reasoning, the twenty-day chart-minus-summary coefficient for the earlier return moves from −8.2 to −3.8 probability percentage points per standard deviation. The corresponding expected-return coefficient moves from −2.04 to −0.68 points (Table G.6). Presentation thus contributes to the differential interpretation of public information emphasized by Kandel and Pearson (1995): the initial form of the history remains reflected in beliefs after both representations are available.

3.5 Interpretation and financial decisions

Reasoning improves investment decisions most when earlier and latest returns have opposite signs. Table 3.5 reports an average expected-payoff gain of 67.4 basis points of claim face value, with a simultaneous 95% interval of [37.9, 96.9]. Both independent requests show the gain, as do the averages for each representation and horizon.

The gain is 113.8 basis points larger when the two returns conflict than when they agree, with an interval of [54.6, 172.9]. It is largest following an earlier rise and a latest-day fall, at 196.0 basis points, compared with 40.8 points following an earlier fall and a latest-day rise. Reasoning strengthens the response to earlier returns and weakens the response to the latest move; its financial benefit is concentrated in histories where those movements point in opposite directions.

A fitted adjustment to probability levels and scales does not reproduce the improvement. The affine comparison in Table 3.5 preserves the ordering of histories by the original no-reasoning probability within each fitted map. Actual reasoning forecasts deliver 81.1 basis points more than the fitted forecasts on held-out histories, with an interval of [50.9, 111.3]. In the first requests, reversals between long and short positions and exits account for most of the payoff improvement; Table G.4 reports the complete accounting.

Table 3.4: Belief convergence and persistence after reciprocal revision

		Reasoning			No reasoning		
Measure	Horizon	Initial	Final	Decline	Initial	Final	Decline
Panel A. Recurring differences and implied positions							
Repeated-gap product (pp^2)	Average	36.3	18.1	18.2	41.1	23.8	17.2
95% interval		[30.8, 41.8]	[14.8, 21.4]	[14.8, 21.6]	[36.7, 45.5]	[21.1, 26.6]	[15.2, 19.2]
Position disagreement (%)	Average	25.47	22.68	2.79	31.49	23.22	8.27
95% interval				[1.3, 4.3]			[6.6, 10.0]
	1 day	32.27	27.90	4.37	36.17	26.27	9.90
	5 days	23.80	21.93	1.87	29.57	21.83	7.73
	20 days	20.33	18.20	2.13	28.73	21.57	7.17
Panel B. Gaps in probability, expected-return and risk forecasts							
Probability gap (pp)	Average	8.26	6.28	1.98	7.03	4.69	2.34
	1 day	4.45	3.59	0.86	4.84	3.31	1.52
	5 days	7.46	5.64	1.81	6.29	4.30	1.98
	20 days	11.37	8.57	2.80	9.25	6.05	3.20
Expected-return gap (pp)	1 day	0.16	0.13	0.02	0.18	0.12	0.06
	5 days	0.68	0.50	0.18	0.82	0.54	0.28
	20 days	2.73	1.78	0.96	3.27	2.02	1.25
Stated-risk gap (pp)	1 day	0.32	0.21	0.11	0.62	0.41	0.21
	5 days	0.69	0.47	0.22	1.32	0.90	0.42
	20 days	1.51	1.07	0.44	2.61	1.85	0.75

Note: Reciprocal revisions start from opposite representations and supply each forecast with the other representation. Each final request retains its own initial forecast and the starting roles and order of the representations. Entries average the three unordered pairs. The initial repeated-gap product is $\mathcal{S}$; the final product is $\mathcal{P}$, the product of the final reciprocal gap and the independent initial gap. Position disagreement is the percentage of comparisons implying different long, flat or short positions at a 100 bp cost; its decline is in percentage points. Forecast gaps are root mean squared differences. Probability gaps are probability percentage points; expected-return and stated-risk gaps are cumulative simple-return percentage points. Average rows average squared gaps, products or disagreement rates across horizons before taking any square root. Declines use unrounded values. Brackets report simultaneous 95% intervals from the eight-contrast disagreement family. Tables D.1 and G.5 give the complete interval family and individual pairs.

Reasoning improves these positions even as average Brier loss rises within every representation (Table 3.1). The more dispersed probabilities select positions with higher expected payoffs despite their greater average squared error. This distinction connects the formation of beliefs to the choices they support: the direction and selection of positions depend on which forecasts cross the investment rule's decision boundaries.

The convergence produced by another representation has a different payoff pattern. With reasoning, expected payoff changes by −6.8 basis points for table receivers, −17.8 for chart receivers and 20.6 for text-summary receivers. The interval for the summary-receiver gain is [−4.5, 45.7]. Without reasoning, all three estimates are negative. Thus, reviewing another representation brings forecasts closer together without producing a uniform improvement in their implied positions. Table G.3 reports the recipient comparisons and probability-loss components, and Table G.1 gives the full family of payoff intervals.[4]

3.6 Second opinions and forecast aggregation

A separately formed opinion offers another interpretation of the history and can improve a combined forecast when the errors offset one another (Bates and Granger, 1969). Table 3.6 separates the contribution of equal averaging from the difference between the final consultation forecast and that average. Comparing peers using the same and different representations holds the mix of constituent forecasts fixed.

With reasoning, almost all of the different-representation advantage in Brier loss comes from averaging: the pooling component is 9.08 of the total 9.12 units on the table's loss scale. Relative to the receiver's initial forecast, however, the final consultation forecast gives back roughly one-quarter of the loss reduction achieved by the average. Without reasoning, final consultation also performs worse than the different-representation average. The independently formed reports improve the combined probability forecast, but the model's subsequent revision reduces that benefit.

[4]Table G.2 reports realized acquisition payoffs at the three evaluated position costs. Table E.1 reports the original loss contrasts and complete-case estimates.

Table 3.5: Reasoning and investment decisions

Sample	First forecast	Independent repeat	Average	Affine adjustment	Actual minus affine
Panel A. Reasoning payoff gains by representation and horizon (bp)					
Numerical table	77.5	74.4	76.0	-15.8	91.8
Chart image	79.5	82.7	81.1	-13.5	94.6
Text summary	34.8	55.6	45.2	-11.7	56.9
All representations and horizons	63.9	70.9	67.4	-13.7	81.1
95% interval			[37.9, 96.9]		[50.9, 111.3]
All representations: 1 day	99.1	114.7	106.9	-35.7	142.6
All representations: 5 days	58.7	64.8	61.7	-15.6	77.3
All representations: 20 days	34.0	33.2	33.6	10.2	23.3
Panel B. Signs of earlier and latest returns					
Earlier rise / latest rise	4.9	15.6	10.2	-22.3	32.5
Earlier fall / latest fall	13.4	21.3	17.4	-4.2	21.6
Earlier rise / latest fall	197.0	194.9	196.0	32.1	163.9
Earlier fall / latest rise	34.5	47.2	40.8	-69.0	109.9
Conflicting minus agreeing	116.3	111.3	113.8	—	—
95% interval			[54.6, 172.9]		

Note: All entries are payoff gains in basis points of claim face value, subtracting no reasoning from reasoning at a 100 bp position cost. Expected payoffs use equation (2.7). The affine adjustment regresses reasoning probabilities on no-reasoning probabilities with an intercept and a nonnegative slope, using five folds and evaluating held-out histories. The fit uses no realized outcomes. Its payoff gain also subtracts the original no-reasoning payoff. The final three columns average the independent requests. Panel B averages representations and horizons; the four groups contain 300, 218, 265 and 212 histories, respectively. Five histories with an exact-zero past return enter Panel A but not Panel B. Brackets show simultaneous 95% intervals from the ten-contrast payoff family. Conflicting minus agreeing compares the weighted means of the two opposite-sign and two same-sign groups. Dashes denote comparisons outside the reported contrast set. Table G.1 reports the full family; Table G.4 reports position-change accounting.

The relative averaging advantage is the direct counterpart of recurring disagreement. Equation (2.8) gives an advantage of $\mathcal{S}/4$ for pooling forecasts from different representations rather than the same representation. Expected-return forecasts exhibit a similar accounting pattern: with reasoning, averaging contributes 0.85 of the 0.86-unit different-representation advantage in normalized return loss. Table H.2 reports both reasoning conditions and all horizons.

Adjusting systematic probability errors sharply reduces the averaging advantage. After separately calibrating the constituent probabilities on training histories, the different-representation advantage falls from 9.08 to 0.40 with reasoning and from 10.27 to 1.34 without it. The calibrated intervals are $[-0.39, 1.19]$ and $[-0.76, 3.43]$, respectively. Much of the original advantage is therefore removed by adjusting probability levels and scales before averaging. Table E.3 reports the complete interval family.[5]

Different-representation consultation raises expected payoff over the initial position by 1.8 basis points with reasoning, with an interval of $[-10.0, 13.6]$, and by 0.1 basis points without reasoning, with an interval of $[-11.5, 11.8]$. Its payoff advantage over same-representation consultation is -2.6 and 2.5 basis points, respectively. Tables H.1 and H.5 report the ordered comparisons and full interval family. Combining opinions improves probability scores through averaging, with little estimated gain in the payoff of the resulting positions.

[5]Table E.2 reports scalar calibration of initial and acquisition forecasts. Table E.4 compares calibration before and after pooling with joint probability fits; Table H.3 reports fitted expected-return comparisons. Table H.4 evaluates the content of probabilities, expected returns and stated risk in held-out predictions.

Table 3.6: Second opinions: averaging, calibration and financial value

	Reasoning			No reasoning		
Peer representation	Pooling gain	Final vs. pool	Total gain	Pooling gain	Final vs. pool	Total gain
Panel A. Original probability forecasts: Brier-loss reduction ($\times 10^4$)						
Same	8.13	-4.23	3.90	0.45	1.48	1.92
Different	17.21	-4.20	13.01	10.71	-1.90	8.81
Different minus same	9.08	0.03	9.12	10.27	-3.38	6.89
Panel B. Calibrated probability forecasts: Brier-loss reduction ($\times 10^4$)						
Same	1.85	-0.60	1.25	0.25	0.11	0.37
Different	2.25	-0.66	1.59	1.59	-1.13	0.46
Different minus same	0.40	-0.07	0.34	1.34	-1.24	0.10
95% interval	[-0.39, 1.19]			[-0.76, 3.43]		
Panel C. Original probability forecasts: expected-payoff gain (bp of claim face value)						
Same	7.4	-3.0	4.3	-4.5	2.1	-2.3
Different	4.6	-2.8	1.8	-0.9	1.1	0.1
95% interval	[-7.3, 16.4]	[-6.6, 1.0]	[-10.0, 13.6]	[-12.5, 10.6]	[-1.5, 3.7]	[-11.5, 11.8]
Different minus same	-2.8	0.2	-2.6	3.5	-1.1	2.5
95% interval			[-10.8, 5.6]			[-7.0, 12.0]

Note: Pooling gain compares the equal average of the receiver and peer probabilities with the receiver's initial probability. Final vs. pool compares the final consultation forecast with that average; total gain compares final with initial. Positive values indicate lower Brier loss in Panels A–B and higher expected payoff in Panel C. Same and different refer to the peer's representation relative to the receiver's. Different minus same subtracts the same-representation gain from the different-representation gain. Panel B separately calibrates the receiver, peer and final probabilities using five-fold held-out evaluation; pooling averages the calibrated constituents. Panels A–B use realized Brier losses multiplied by 10^4. Panel C uses original probabilities and conditional expected payoffs at a 100 bp position cost. All entries average horizons and the relevant ordered pairs; totals use unrounded components. Panel B intervals apply to the different-minus-same pooling gain and are simultaneous within the eight calibrated probability-loss contrasts in Table E.3. Panel C intervals are simultaneous within the eight consultation-payoff contrasts in Table H.5. Table E.4 reports calibrated loss levels and Table H.1 gives conditional losses and payoffs by ordered pair.

4 Conclusion

The presentation of financial information produces persistent differences in the assessments supplied by a common large language model. In a controlled experiment, numerical tables, chart images and text summaries of the same market histories generate different forecasts of returns and risk. Independent repetitions establish that the probability differences recur across requests. Reviewing another representation brings forecasts closer together, but the starting presentation remains reflected in the revised assessment. Even after both representations are available, forecasts imply different investment positions in about 23% of comparisons under a common decision rule. Access to the same market inputs therefore reduces disagreement without eliminating the influence of the initial presentation.

These differences reflect distinct responses to past returns. Text summaries produce the strongest extrapolation from earlier returns, both in the probability of a gain and in the expected return. Enabling the model's reasoning setting increases the response to earlier returns and reduces the response to the latest daily move. The resulting positions have higher expected payoffs, with the largest improvement when earlier and latest returns point in opposite directions. This improvement occurs even as average squared probability error rises. How the history is interpreted matters for the positions selected, and changes in forecast accuracy do not fully describe those financial consequences.

A separately formed opinion produces a different pattern. Averaging forecasts based on different representations improves probability accuracy relative to averaging forecasts from the same representation. Most of this advantage disappears after adjusting systematic probability errors, and the model's subsequent revision gives back part of the improvement achieved by averaging. The estimated payoff gains from consultation are small. Together, the revision results distinguish three outcomes of financial analysis: agreement between assessments, accuracy of forecasts and expected payoff of implied positions. Reviewing the underlying information brings beliefs closer together; combining opinions improves probability scores; more intensive reasoning improves the decisions supported by the

forecasts. These benefits need not coincide.

The findings identify presentation as a source of financial disagreement that survives review of the underlying evidence. Investors using the same analytical model can receive different assessments even after supplying the same market representations, because the starting assessment remains relevant to the final forecast. Understanding the financial beliefs supplied by AI consequently requires attention to how information is presented and how assessments are revised. A common model does not by itself provide a common interpretation, and the value of further analysis depends on the investment decisions it changes.

References

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Ebens, H. (2001). The distribution of realized stock return volatility. *Journal of Financial Economics*, 61(1):43–76.

Antweiler, W. and Frank, M. Z. (2004). Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. *Journal of Finance*, 59(3):1259–1294.

Banerjee, S. (2011). Learning from Prices and the Dispersion in Beliefs. *Review of Financial Studies*, 24(9):3025–3068.

Banerjee, S. and Kremer, I. (2010). Disagreement and Learning: Dynamic Patterns of Trade. *Journal of Finance*, 65(4):1269–1302.

Bates, J. M. and Granger, C. W. J. (1969). The Combination of Forecasts. *Operational Research Quarterly*, 20(4):451–468.

Beshears, J., Choi, J. J., Laibson, D., and Madrian, B. C. (2017). Does Aggregated Returns Disclosure Increase Portfolio Risk Taking? *Review of Financial Studies*, 30(6):1971–2005.

Bhagwat, V., Cookson, J. A., Dim, C., and Niessner, M. (2026). The Market's Mirror: Revealing Investor Disagreement with LLMs. Working paper.

Bini, P., Cong, L. W., Huang, X., and Jin, L. J. (2026). Behavioral Economics of AI: LLM Biases and Corrections. Working paper.

Blackwell, D. (1953). Equivalent Comparisons of Experiments. *Annals of Mathematical Statistics*, 24(2):265–272.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

Bybee, J. L. (2025). The Ghost in the Machine: Generating Beliefs with Large Language Models. Working paper, University of Chicago.

Cao, S., Jiang, W., Wang, J., and Yang, B. (2024). From Man vs. Machine to Man

+ Machine: The Art and AI of Stock Analyses. *Journal of Financial Economics*, 160:103910.

Cao, S., Jiang, W., Yang, B., and Zhang, A. L. (2023). How to Talk When a Machine Is Listening: Corporate Disclosure in the Age of AI. *Review of Financial Studies*, 36(9):3603–3642.

Cassella, S. and Gulen, H. (2018). Extrapolation Bias and the Predictability of Stock Returns by Price-Scaled Variables. *Review of Financial Studies*, 31(11):4345–4397.

Chen, H., De, P., Hu, Y. J., and Hwang, B.-H. (2014). Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media. *Review of Financial Studies*, 27(5):1367–1403.

Chen, S., Green, T. C., Gulen, H., and Zhou, D. (2025). What Does ChatGPT Make of Historical Stock Returns? Extrapolation and Miscalibration in LLM Stock Return Forecasts. Working paper, December.

Cookson, J. A., Engelberg, J. E., and Mullins, W. (2023). Echo Chambers. *Review of Financial Studies*, 36(2):450–500.

Cookson, J. A. and Niessner, M. (2020). Why Don't We Agree? Evidence from a Social Network of Investors. *Journal of Finance*, 75(1):173–228.

Da, Z., Huang, X., and Jin, L. J. (2021). Extrapolative beliefs in the cross-section: What can we learn from the crowds? *Journal of Financial Economics*, 140(1):175–196.

D'Acunto, F., Prabhala, N., and Rossi, A. G. (2019). The Promises and Pitfalls of Robo-Advising. *Review of Financial Studies*, 32(5):1983–2020.

De Bondt, W. F. M. and Thaler, R. (1985). Does the Stock Market Overreact? *Journal of Finance*, 40(3):793–805.

Diamond, D. W. (1985). Optimal Release of Information by Firms. *Journal of Finance*, 40(4):1071–1094.

Diamond, D. W. and Verrecchia, R. E. (1981). Information Aggregation in a Noisy Rational Expectations Economy. *Journal of Financial Economics*, 9(3):221–235.

Enke, B. and Zimmermann, F. (2019). Correlation Neglect in Belief Formation. *Review of Economic Studies*, 86(1):313–332.

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *Journal of Finance*, 25(2):383–417.

Fedyk, A. (2024). Front-Page News: The Effect of News Positioning on Financial Markets. *Journal of Finance*, 79(1):5–33.

Giglio, S., Maggiori, M., Stroebel, J., and Utkus, S. (2021). Five Facts about Beliefs and Portfolios. *American Economic Review*, 111(5):1481–1522.

Glaser, M., Iliewa, Z., and Weber, M. (2019). Thinking about Prices versus Thinking about Returns in Financial Markets. *Journal of Finance*, 74(6):2997–3039.

Gneiting, T. and Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Goldstein, I. and Yang, L. (2015). Information Diversity and Complementarities in Trading and Information Acquisition. *Journal of Finance*, 70(4):1723–1765.

Greenwood, R. and Shleifer, A. (2014). Expectations of Returns and Expected Returns. *Review of Financial Studies*, 27(3):714–746.

Grossman, S. J. and Stiglitz, J. E. (1980). On the Impossibility of Informationally Efficient Markets. *American Economic Review*, 70(3):393–408.

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. *Review of Financial Studies*, 33(5):2223–2273.

Halim, E., Riyanto, Y. E., and Roy, N. (2019). Costly Information Acquisition, Social Networks, and Asset Prices: Experimental Evidence. *Journal of Finance*, 74(4):1975–2010.

Hamilton, J. D. (1989). A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. *Econometrica*, 57(2):357–384.

Harris, M. and Raviv, A. (1993). Differences of Opinion Make a Horse Race. *Review of Financial Studies*, 6(3):473–506.

Hartzmark, S. M. and Shue, K. (2018). A Tough Act to Follow: Contrast Effects in Financial Markets. *Journal of Finance*, 73(4):1567–1613.

Hellwig, M. F. (1980). On the Aggregation of Information in Competitive Markets. *Journal of Economic Theory*, 22(3):477–498.

Hirshleifer, D. and Teoh, S. H. (2003). Limited attention, information disclosure, and financial reporting. *Journal of Accounting and Economics*, 36(1-3):337–386.

Hong, H., Kubik, J. D., and Stein, J. C. (2004). Social Interaction and Stock-Market Participation. *Journal of Finance*, 59(1):137–163.

Jegadeesh, N. (1990). Evidence of Predictable Behavior of Security Returns. *Journal of Finance*, 45(3):881–898.

Jegadeesh, N. and Titman, S. (1993). Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency. *Journal of Finance*, 48(1):65–91.

Jiang, J., Kelly, B., and Xiu, D. (2023). (Re-)Imag(in)ing Price Trends. *Journal of Finance*, 78(6):3193–3249.

Kandel, E. and Pearson, N. D. (1995). Differential Interpretation of Public Signals and Trade in Speculative Markets. *Journal of Political Economy*, 103(4):831–872.

Kim, A. G., Muhn, M., and Nikolaev, V. V. (2024). Financial Statement Analysis with Large Language Models. Working paper, November version.

Lopez-Lira, A. and Tang, Y. (2026). Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models. *Journal of Financial Economics*, 184:104335.

Merton, R. C. (1976). Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics*, 3(1-2):125–144.

Mondria, J., Vives, X., and Yang, L. (2022). Costly Interpretation of Asset Prices. *Management Science*, 68(1):52–74.

Peng, L. and Xiong, W. (2006). Investor attention, overconfidence and category learning. *Journal of Financial Economics*, 80(3):563–602.

Pool, V. K., Stoffman, N., and Yonker, S. E. (2015). The People in Your Neighborhood: Social Interactions and Mutual Fund Portfolios. *Journal of Finance*, 70(6):2679–2732.

Schwert, G. W. (1990). Stock Volatility and the Crash of '87. *Review of Financial Studies*, 3(1):77–102.

Van Nieuwerburgh, S. and Veldkamp, L. (2009). Information Immobility and the Home Bias Puzzle. *Journal of Finance*, 64(3):1187–1215.

Appendix

A Derivations

A.1 Information and reported probabilities

Let $D \in \{0, 1\}$ be the directional outcome, X the complete rounded history, $X_j = g_j(X)$ the representation and Z_j the report containing p_j. Processing randomness is independent of the outcome conditional on the input, so $D - X - X_j - Z_j$ is a Markov chain. Define $\mu_X = \mathbb{E}[D \mid X]$, $\mu_j = \mathbb{E}[D \mid X_j]$ and $\mu_{Z_j} = \mathbb{E}[D \mid Z_j]$. The loss decomposition is

$$\begin{aligned}\mathbb{E}[(D - p_j)^2] = \mathbb{E}[(D - \mu_X)^2] + \mathbb{E}[(\mu_X - \mu_j)^2] \\ + \mathbb{E}[(\mu_j - \mu_{Z_j})^2] + \mathbb{E}[(\mu_{Z_j} - p_j)^2].\end{aligned} \tag{A.1.1}$$

The four terms are residual uncertainty, representation loss, processing loss and calibration error.

The chain implies $\mathbb{E}[D \mid X, X_j, Z_j] = \mu_X$, $\mathbb{E}[\mu_X \mid X_j, Z_j] = \mu_j$ and $\mathbb{E}[\mu_j \mid Z_j] = \mu_{Z_j}$. Expand

$$D - p_j = (D - \mu_X) + (\mu_X - \mu_j) + (\mu_j - \mu_{Z_j}) + (\mu_{Z_j} - p_j).$$

Successive conditioning sets every cross-product to zero, giving equation (A.1.1). The outcome, reported probability and conditional probabilities are bounded.

A.2 Repeated representation differences

Write $p_j^{(b)} = b_j(X) + \epsilon_j^{(b)}$, where $\mathbb{E}[\epsilon_j^{(b)} \mid X] = 0$. Conditional independence of the two repetitions gives

$$\mathbb{E}[(p_j^{(1)} - p_k^{(1)})(p_j^{(2)} - p_k^{(2)}) \mid X] = (b_j(X) - b_k(X))^2.$$

Taking expectations proves equation (2.4).

For each episode and horizon, write $x_j = p_j^{(1)}$, $y_j = p_j^{(2)}$, with representation means $\bar{x}$ and $\bar{y}$. Then

$$s = \frac{1}{3}\sum_{j<k}(x_j - x_k)(y_j - y_k) = \sum_{j=1}^{3}(x_j - \bar{x})(y_j - \bar{y}).$$

The across-format and within-format squared gaps satisfy

$$\frac{1}{6}\sum_{j\neq k}(x_j - y_k)^2 - \frac{1}{3}\sum_{j}(x_j - y_j)^2 = s.$$

Averaging gives $\mathcal{S} = \langle s \rangle$ without an independence assumption.

A.3 Balanced pooling

Equation (2.8) follows by expansion. Balanced receiver and peer frequencies equate constituent losses across same-format and different-format pairs, giving

$$L^{\text{pool}}_{\text{same}} - L^{\text{pool}}_{\text{different}} = \frac{\mathcal{S}}{4}.$$

This holds for realized and conditional losses. For final forecast p_f and initial forecast p_i,

$$L(p_i) - L(p_f) = \{L(p_i) - L(\bar{p})\} + \{L(\bar{p}) - L(p_f)\}.$$

The terms are the pooling and revision components.

B Generating process and calibration

B.1 State transitions, variance and jumps

The state ordering is continuation (M), reversal (R), no conditional log-return predictability (N), and stress (S). The transition matrix is

$$A = \begin{pmatrix} .950 & .014 & .024 & .012 \\ .014 & .950 & .024 & .012 \\ .015 & .015 & .960 & .010 \\ .030 & .030 & .040 & .900 \end{pmatrix}, \qquad A_{ab} = \Pr(s_t = b \mid s_{t-1} = a).$$

Its stationary distribution is $\boldsymbol{\omega} = (.25, .25, .40, .10)$, with $\boldsymbol{\omega} A = \boldsymbol{\omega}$. In equations (2.1)–(2.2), variance and jumps are

$$v_t = \bar{\sigma}_{s_t}^2 \exp(\tau_v \nu_t - \tau_v^2/2), \qquad \nu_t \sim N(0,1), \tag{B.1.1}$$

$$B_t \mid s_t = s \sim \text{Bernoulli}(\lambda_s), \qquad J_t = B_t(\mu_s^J + \sigma_s^J \zeta_t), \quad \zeta_t \sim N(0,1). \tag{B.1.2}$$

Variance is drawn independently each date conditional on the state; persistence arises through the state process. In stress, $\delta_S = -\lambda_S \mu_S^J$. The state-conditional innovation variance, integrating over variance and jump draws, is

$$\mathcal{V}_s = \bar{\sigma}_s^2 + \lambda_s (\sigma_s^J)^2 + \lambda_s (1 - \lambda_s)(\mu_s^J)^2.$$

Table B.1 reports the return parameters.

Table B.1: Return-process parameters

Parameter	Continuation	Reversal	No predictability	Stress
Drift offset δ_s (bp)	0.00	0.00	0.00	47.32
Latest-return coefficient ϕ_s	0.234	-0.370	0.000	0.000
Five-day coefficient $\beta_{5,s}$	0.170	-0.268	0.000	0.000
Twenty-day coefficient $\beta_{20,s}$	0.314	-0.057	0.000	0.000
Diffusion scale $\bar{\sigma}_s$ (%)	1.207	1.126	1.142	1.892
Jump probability λ_s (%)	0.803	0.803	0.803	8.422
Jump mean μ_s^J (%)	0.000	0.000	0.000	-5.619
Jump SD σ_s^J (%)	0.642	0.642	0.642	3.012

Note: Parameters are rounded for display. The common daily log-return drift is $\mu_0 = 4.3823$ bp and within-state log-variance dispersion is $\tau_v = 0.6829$. The stress drift offset satisfies $\delta_S = -\lambda_S \mu_S^J$ at full precision. Diffusion and jump scales are in daily log-return units.

B.2 Completed daily observations

The opening gap and opening price are

$$g_t = a_g r_t + d_g J_t + \sigma_g \sqrt{v_t}\xi_t, \qquad O_t = C_{t-1} e^{g_t}, \qquad \xi_t \sim N(0,1).$$

Let $b_t = |r_t - g_t|$ denote the absolute log body and set

$$\omega_t = \alpha_w \sqrt{v_t} + \gamma_w b_t + \chi_w |J_t|.$$

The upper and lower log wicks are conditionally independent gamma draws with shape κ and scale ω_t/κ. The high and low are

$$H_t = \max(O_t, C_t) e^{w_t^+}, \qquad L_t = \min(O_t, C_t) e^{-w_t^-}.$$

Log volume satisfies

$$\begin{aligned} \log V_t = {} & \mu_V + \rho_V(\log V_{t-1} - \mu_V) + b_v \log(v_t/\bar{v}_0) \\ & + b_b b_t/\sqrt{v_t} + b_J B_t + \sigma_V \eta_t, \qquad \eta_t \sim N(0,1), \end{aligned} \tag{B.2.1}$$

where $\bar{v}_0 = (\omega_M \bar{\sigma}_M^2 + \omega_R \bar{\sigma}_R^2 + \omega_N \bar{\sigma}_N^2)/(\omega_M + \omega_R + \omega_N)$. Gaussian primitive innovations are mutually independent across components and time and independent of state transitions and jump occurrence. Jump indicators are independent over time conditional on the state path. Primitive wick draws are independent of each other and of all other innovations. Table B.2 gives the observation parameters.

Initialization draws the state from $\boldsymbol{\omega}$, sets twenty drift-adjusted return lags to zero, and sets log volume to μ_V. A 500-date burn-in precedes the retained history. The close is normalized to 100 at the start of the retained sixty-date window. Only the final twenty dates are displayed.

Table B.2: Completed-bar and volume parameters

Parameter	Symbol	Value
Return loading in opening gap	a_g	0.2477
Jump loading in opening gap	d_g	0.0758
Opening-gap noise scale	σ_g	0.4734
Variance loading in wick mean	α_w	0.2716
Body loading in wick mean	γ_w	0.6124
Jump loading in wick mean	χ_w	0.7507
Gamma wick shape	κ	7.4451
Log-volume persistence	ρ_V	0.6530
Variance loading in volume	b_v	0.3532
Body loading in volume	b_b	0.4724
Jump loading in volume	b_J	0.6496
Log-volume noise scale	σ_V	0.5851
Log-volume location	μ_V	$\log(10^6)$
Non-stress variance reference	$\bar{v}_0$	1.3364×10^{-4}

Note: The parameters enter the observation equations in Appendix B. Values are rounded from the earlier source specification. The variance reference is the stationary mean diffusion variance conditional on a non-stress state.

B.3 Recorded validation statistics

Parameters were fixed before the forecasting experiment; AI outputs were not used in calibration. After selecting the return process, the completed-bar and volume blocks were selected from bounded parameter grids. Each block evaluated 512 quasi-random vectors and one reference vector, using common primitive draws. Screening used 50,000 retained dates, followed by 500,000 dates for 32 candidates and two million dates for eight

candidates. Selection minimized the sum of squared deviations from target midpoints, scaled by target half-widths, among vectors satisfying all restrictions.

Table B.3 reports the recorded final validation statistics from five independent replications of one million retained dates each. These replications precede and are separate from the 1,000 forecasting episodes. Daily returns in this table are log returns. The predictive-component variance share is $\mathrm{Var}(m_t(s) \mid s_t = s)/\mathrm{Var}(r_t - \mu_0 \mid s_t = s)$. The stress variance ratio is $\mathcal{V}_S/\max_{s \neq S} \mathcal{V}_s$. The variance-state AUC uses latent daily variance to distinguish stress from non-stress observations.

Table B.3: Validation statistics of the financial environment

Statistic	Unit	Recorded value
Mean daily log return	bp	4.53
Positive-return frequency	%	52.35
Daily log-return standard deviation	%	1.42
Jump frequency	%	1.58
Log-return skewness		-0.809
Log-return excess kurtosis		7.397
Maximum absolute return autocorrelation, lags 1–5		0.066
Mean absolute return autocorrelation, lags 1–5		0.029
Absolute-return autocorrelation, lag 1		0.125
Mean absolute-return autocorrelation, lags 1–20		0.035
Predictive-component variance share: continuation		0.151
Predictive-component variance share: reversal		0.181
Stress innovation-variance ratio		4.642
Variance-state AUC		0.846
Median daily price range / previous close	%	2.06
Opening-gap variance share		0.241
Correlation: log volume and absolute log return		0.372
Log-volume autocorrelation, lag 1		0.702
Standardized stress-volume difference		1.083

Note: The table reproduces the validation estimates recorded before the forecasting experiment, based on five independent paths of one million retained dates each. The predictive shares, stress variance ratio and variance-state AUC are defined in Appendix B. Return moments refer to daily log returns. The table does not use AI forecasts or the 1,000 forecasting episodes.

C Representation construction

For each displayed price $P_t \in \{O_t, H_t, L_t, C_t\}$ and date $t = -19, \ldots, 0$, the numerical input is

$$\widetilde{P}_t = \mathrm{round}_2(100 P_t / C_{-19}), \qquad \widetilde{V}_t = \mathrm{round}_3(V_t / 10^6).$$

Rounding is to nearest with ties to even. The table contains all five fields in chronological order. Figure 2.1 reproduces the matched input example. All chart coordinates and summary statistics use this rounded panel.

The chart is a 768-by-512-pixel candlestick-and-volume image. Price and volume occupy separate panels with a four-to-one height ratio. Rising and falling candles have different colors, and the final close is marked on the right-hand price scale. Volume is represented by bar heights without printed numerical values. The chart contains no moving average, technical indicator, state estimate or forecast.

The summary uses a fixed sequence of fields: number of dates; initial and final closing indices and cumulative percentage change; latest close-to-close percentage change; count of positive close-to-close changes; daily close-return volatility; largest and smallest daily changes; displayed price minimum and maximum; mean daily high-to-low range; latest and mean volume; and latest volume divided by mean volume. Close-to-close changes are computed as $\widetilde{C}_t / \widetilde{C}_{t-1} - 1$. Wording, field order and numerical precision are fixed across histories. The summary includes no autocorrelation estimate, technical indicator, state label or interpretive statement.

The final displayed bar is completed before the forecast is requested. Forecast horizons begin on the next date. Latent states, variance draws, jump realizations, structural parameters and future observations are excluded from all three representations and all forecast requests.

D Measurement and inference

State weights are 0.25, 0.25, 0.40 and 0.10 for continuation, reversal, no predictability and stress. They equal the corresponding fractions of the 1,000-episode sample. All averages retain episode pairing. Horizon averages are arithmetic means of separately evaluated horizon-specific quantities.

D.1 Forecasts and disagreement

Let $p^{(1)}_{ejh}$ and $p^{(2)}_{ejh}$ denote first and independent-repeat probabilities for episode e, representation j and horizon h. For an unordered pair (j, k), define

$$d^{(b)}_{ejkh} = p^{(b)}_{ejh} - p^{(b)}_{ekh}, \quad b \in \{1, 2\}, \tag{D.1.1}$$

$$d^{(+)}_{ejkh} = p^{(+)}_{ej\to k,h} - p^{(+)}_{ek\to j,h}. \tag{D.1.2}$$

Pair orientation is raw minus chart, raw minus summary and chart minus summary. With $\langle\cdot\rangle$ denoting the prescribed average over episodes, pairs and horizons, the initial and final products are

$$\mathcal{S} = \langle d^{(1)} d^{(2)} \rangle, \qquad \mathcal{P} = \langle d^{(+)} d^{(2)} \rangle.$$

Their difference is $\mathcal{S} - \mathcal{P}$. The across-format minus within-format squared-gap identity is derived in Appendix A. Probability gaps are multiplied by 100 and gap products by 10^4. RMS values take the square root after averaging squared gaps. Table D.1 reports all eight interval estimates.

Acquisition requests contain the receiver's initial forecast, its original representation and the added representation. They contain no forecast produced from the added representation. Reciprocal routes contain the same pair of representation sources in different starting roles and presentation orders. Raw-involving routes contain the complete rounded panel. Consultation requests contain the receiver's initial forecast and a peer forecast from the independent repeat. The three same-representation and six different-representation

consultation pairs have balanced receiver and peer representation frequencies.

D.2 Financial positions and payoffs

Equation (2.6) uses prices Q_h of 0.522971, 0.539136 and 0.565338 at one, five and twenty days. These prices equal positive-return frequencies in a separate pre-experiment simulation of one million episodes. The main cost is $k = 0.01$. Expected payoffs follow equation (2.7); realized payoffs replace q_{eh} by D_{eh}. Multiplication by 10^4 expresses payoff differences in basis points of claim face value. Each horizon's position is chosen before averaging across horizons.

The stored conditional probabilities use 8,192 future continuations per episode, split between two independent streams of 4,096. Future paths are shared across horizons within a continuation. Conditioning includes the origin's latent regime and twenty adjusted return lags. These variables and the resulting probabilities are used by the evaluator, not supplied to the model or used to choose positions. Numerical standard errors use the continuation covariance across horizons and are reported separately from episode-bootstrap uncertainty.

D.3 Loss accounting and estimation

The Brier loss is $(D - p)^2$. Its conditional expectation is $p^2 + q(1 - 2p)$. For an initial probability p_i and a final probability p_f, the loss reduction is

$$\begin{aligned} L_i - L_f = {} & (\bar{D} - \bar{p}_i)^2 - (\bar{D} - \bar{p}_f)^2 \\ & + \mathrm{Var}(p_i) - \mathrm{Var}(p_f) + 2\{\mathrm{Cov}(D, p_f) - \mathrm{Cov}(D, p_i)\}. \end{aligned} \tag{D.3.1}$$

All moments use the prescribed episode weights. The equal-pool decomposition and balanced-pair identity are given in Appendix A.

Earlier returns are $\log(C_{-1}/C_{-19})$ and the latest return is $\log(C_0/C_{-1})$, using rounded displayed closes. Their global means are 0.00879 and 0.00078, and their global standard

deviations are 0.07003 and 0.01424. Regressions include an intercept and both standardized returns. The separate full-history specification uses $\log(C_0/C_{-19})$ and the latest return, with the standardization stored for that specification.

All fitted mappings use five episode folds, each containing 200 held-out histories. Each fold contains 50, 50, 80 and 20 histories from the four states. Forecasts, representations, horizons and reasoning conditions for an episode remain in the same fold. Scalar probability calibration regresses the realized binary outcome on an intercept and the reported probability. The reasoning adjustment instead regresses the reasoning probability on the no-reasoning probability, with an intercept and a nonnegative slope. Its target contains no outcomes or evaluator probabilities. Predictions are clipped to $[0, 1]$ when required.

The original contrast families use 9,999 paired state-stratified episode draws. Subsequent uncertainty families use 4,999 such draws. Every draw retains the original state counts and carries all observations for a sampled episode together. Intervals use the full-sample estimate plus or minus the family-specific critical value times the bootstrap standard deviation. The critical value is the 95th percentile of the maximum absolute standardized deviation within the stated family. Bootstrap duplicates retain their original fold. The required calibration or probability-adjustment mappings are refitted using sampled training observations. Subsequent families were specified after inspection of the original results. Their simultaneous coverage applies within each reported family.

No initial forecast required a fallback. Acquisition contains four fallback forecasts among 6,000 scientific observations with reasoning and seven without reasoning. Consultation contains twenty-one among 9,000 scientific observations with reasoning and eleven without reasoning. Copying controls are separate observations. The recorded no-copy counts are 154 per reasoning condition for acquisition and 229 and 155 for consultation. Complete-case estimates retain 971 histories.

Table D.1: Disagreement and reciprocal acquisition: interval estimates

Contrast	Unit	Estimate	95% interval	Bootstrap SD
Reasoning				
Initial repeat product	pp^2	36.33	[30.83, 41.84]	2.10
Final repeat product	pp^2	18.09	[14.78, 21.40]	1.26
Reduction in repeat product	pp^2	18.24	[14.83, 21.65]	1.30
Reduction in action disagreement	pp	2.79	[1.32, 4.26]	0.56
No reasoning				
Initial repeat product	pp^2	41.06	[36.67, 45.45]	1.67
Final repeat product	pp^2	23.85	[21.09, 26.60]	1.05
Reduction in repeat product	pp^2	17.21	[15.22, 19.21]	0.76
Reduction in action disagreement	pp	8.27	[6.56, 9.97]	0.65

Note: Initial repeat products multiply representation gaps across the two independent initial forecasts. Final repeat products multiply the reciprocal-acquisition gap by the independent initial gap. Reduction subtracts the final from the initial product. Action disagreement is the share of reciprocal routes implying different long, flat or short positions at a 100 bp position cost. Entries average the three representation pairs and the three horizons. Intervals are simultaneous over all eight rows and use 4,999 paired state-stratified episode draws. pp denotes percentage points.

Table D.2: Initial disagreement by representation pair

Pair	Horizon	RMS gap (pp)	Repeated-gap product (pp^2)	Position disagreement (%)
Reasoning				
Raw–chart	1 day	3.81	3.48	35.60
	5 days	6.17	10.58	27.60
	20 days	9.27	25.49	24.70
	Average	6.80	13.18	29.30
Raw–summary	1 day	4.55	10.17	25.70
	5 days	7.61	31.04	16.80
	20 days	11.53	70.58	14.40
	Average	8.40	37.26	18.97
Chart–summary	1 day	4.92	15.36	35.50
	5 days	8.42	47.53	27.00
	20 days	13.01	112.77	21.90
	Average	9.38	58.56	28.13
No reasoning				
Raw–chart	1 day	2.78	5.51	31.70
	5 days	3.61	8.47	27.40
	20 days	5.79	21.86	25.90
	Average	4.25	11.95	28.33
Raw–summary	1 day	5.12	22.63	34.30
	5 days	6.93	40.89	28.20
	20 days	10.20	85.62	27.70
	Average	7.71	49.72	30.07
Chart–summary	1 day	6.02	31.90	42.50
	5 days	7.58	49.82	33.10
	20 days	10.92	102.84	32.60
	Average	8.42	61.52	36.07

Note: RMS gaps and position disagreement use first forecasts. Repeated-gap products multiply each signed representation gap by the corresponding gap in the independent repeat. Positions follow equation (2.6) at a 100 bp cost. Horizon averages first average squared gaps, products or disagreement rates; RMS then takes the square root.

E Original contrasts and probability mappings

Table E.1: Original contrasts and complete-case estimates

Contrast	All histories	95% interval	Complete cases	95% interval
Panel A. Return-loss contrasts				
Acquisition: return loss	2.12	[1.02, 3.22]	2.08	[0.96, 3.19]
Consultation diversity: return loss	0.86	[0.42, 1.29]	0.79	[0.35, 1.24]
Panel B. Probability-loss contrasts				
Acquisition: Brier loss	21.21	[9.77, 32.65]	19.60	[8.01, 31.20]
Consultation diversity: Brier loss	9.12	[4.65, 13.58]	8.35	[4.00, 12.71]
Panel C. Reasoning interactions in return loss				
Reasoning interaction: acquisition	0.26	[-0.80, 1.31]	0.07	[-1.00, 1.15]
Reasoning interaction: consultation	-1.22	[-1.70, -0.74]	-1.31	[-1.80, -0.82]
Panel D. Standalone return-loss differences				
Raw minus chart	1.47	[-1.11, 4.05]	1.54	[-1.05, 4.14]
Raw minus summary	-5.36	[-9.17, -1.54]	-5.06	[-8.99, -1.13]
Chart minus summary	-6.83	[-11.02, -2.63]	-6.60	[-10.86, -2.34]

Note: All-history estimates use 1,000 episodes and the original fallback rules. Complete-case estimates use 971 episodes. Return-loss entries equal the recorded normalized squared-return-loss differences multiplied by 100; Brier entries are multiplied by 10^4. Return loss is divided by the squared fixed reference scale for the corresponding horizon. Positive acquisition and consultation-diversity entries denote loss reductions. Reasoning interactions subtract no-reasoning estimates. Representation comparisons subtract the second loss from the first. The original analysis uses 9,999 paired state-stratified bootstrap draws; intervals are simultaneous within each panel. The acquisition and consultation levels in Panels A and B use reasoning.

Table E.2: Scalar probability calibration

Reasoning	Horizon	Initial: original	Initial: calibrated	Acquired: original	Acquired: calibrated
Reasoning	1 day	2,509.28	2,495.85	2,501.88	2,492.67
	5 days	2,512.94	2,434.34	2,505.61	2,434.05
	20 days	2,626.59	2,434.96	2,577.68	2,433.72
	Average	2,549.60	2,455.05	2,528.39	2,453.48
No reasoning	1 day	2,515.22	2,501.48	2,520.30	2,500.46
	5 days	2,502.16	2,438.02	2,506.40	2,437.88
	20 days	2,515.74	2,427.99	2,503.35	2,425.12
	Average	2,511.04	2,455.83	2,510.01	2,454.49

Note: Entries are Brier losses multiplied by 10^4. Each scalar map regresses the binary outcome on an intercept and one reported probability using four of five folds and evaluates the held-out fold. Initial forecasts refer to the first forecast. Acquisition averages the six ordered representation pairs.

Table E.3: Calibrated probability-loss contrasts

Reasoning	Contrast	Estimate	95% interval
Reasoning	Frequency forecast minus different-format pool	12.32	[-16.58, 41.22]
Reasoning	Same-format minus different-format pool	0.40	[-0.39, 1.19]
Reasoning	Consultation final minus different-format pool	0.66	[-2.07, 3.40]
Reasoning	Initial minus acquisition	1.57	[-3.83, 6.96]
No reasoning	Frequency forecast minus different-format pool	10.87	[-14.42, 36.17]
No reasoning	Same-format minus different-format pool	1.34	[-0.76, 3.43]
No reasoning	Consultation final minus different-format pool	1.13	[-2.51, 4.76]
No reasoning	Initial minus acquisition	1.34	[-3.54, 6.23]

Note: Entries are Brier-loss differences multiplied by 10^4, averaged equally across horizons. Constituent probabilities are calibrated separately before equal pooling. The frequency forecast uses the event frequency in the training sample. Intervals are simultaneous over the eight contrasts, with calibration maps refitted in each of 4,999 paired state-stratified episode draws.

Table E.4: Two-probability mappings and consultation

Representations	Horizon	Initial	Separate calibration, pool	Pool, then calibration	Final calibration	Joint fit
Reasoning						
Same	1 day	2,495.85	2,494.29	2,494.30	2,495.16	2,500.53
Same	5 days	2,434.34	2,432.38	2,432.55	2,432.41	2,438.21
Same	20 days	2,434.96	2,432.92	2,432.85	2,433.81	2,434.63
Same	Average	2,455.05	2,453.19	2,453.23	2,453.79	2,457.79
Different	1 day	2,495.85	2,493.96	2,494.61	2,495.21	2,498.56
Different	5 days	2,434.34	2,431.91	2,432.06	2,432.39	2,435.18
Different	20 days	2,434.96	2,432.50	2,432.61	2,432.77	2,436.10
Different	Average	2,455.05	2,452.79	2,453.09	2,453.46	2,456.62
No reasoning						
Same	1 day	2,501.48	2,500.87	2,500.98	2,500.83	2,505.53
Same	5 days	2,438.02	2,438.23	2,438.23	2,438.07	2,442.72
Same	20 days	2,427.99	2,427.62	2,427.57	2,427.48	2,430.05
Same	Average	2,455.83	2,455.58	2,455.59	2,455.46	2,459.43
Different	1 day	2,501.48	2,500.31	2,500.33	2,500.55	2,508.21
Different	5 days	2,438.02	2,437.10	2,436.83	2,437.80	2,440.02
Different	20 days	2,427.99	2,425.31	2,427.25	2,427.75	2,423.48
Different	Average	2,455.83	2,454.24	2,454.81	2,455.37	2,457.24

Note: Entries are realized Brier losses multiplied by 10^4. The joint fit regresses the binary outcome on an intercept and the receiver and peer probabilities using four training folds, then evaluates the held-out fold. All procedures use the same five folds. Separate calibration fits each constituent probability before pooling; pool calibration fits the uncalibrated arithmetic mean. Same and different representations average three and six ordered pairs, respectively.

F Past returns, expected returns and stated risk

Table F.1: Past returns and probability forecasts in independent repeats

Representation	Forecast	1 day: earlier	1 day: latest	5 days: earlier	5 days: latest	20 days: earlier	20 days: latest
Reasoning							
Raw table	First	4.32	1.19	8.07	2.53	11.92	3.66
Raw table	Repeat	4.37	1.10	8.35	2.39	12.16	3.45
Chart	First	3.69	0.98	6.84	2.03	9.97	2.94
Chart	Repeat	3.64	1.01	6.99	2.05	10.33	2.99
Summary	First	6.64	1.57	12.29	2.91	18.14	4.22
Summary	Repeat	6.70	1.46	12.47	3.00	18.53	4.55
No reasoning							
Raw table	First	1.40	2.40	3.05	3.24	5.04	4.74
Raw table	Repeat	1.46	2.37	3.10	3.19	5.11	4.67
Chart	First	0.89	1.78	2.72	2.29	4.61	3.25
Chart	Repeat	0.94	1.74	2.75	2.30	4.66	3.34
Summary	First	4.99	2.29	7.78	3.06	11.50	4.41
Summary	Repeat	4.89	2.39	7.75	3.10	11.49	4.45

Note: Entries are coefficients from separate regressions of stated probabilities on an intercept, the earlier 18-interval log return and the latest one-day log return. Both returns are standardized using their full-sample standard deviations. Units are probability percentage points per standard deviation. The first and repeat forecasts are independent calls on the same representation and history.

Table F.2: Stated expected returns and risk

Representation	Horizon	Mean return	Mean SD	Return: earlier	Return: latest	SD: earlier	SD: latest
Reasoning							
Raw table	1 day	0.03	1.42	0.17	0.04	-0.03	-0.07
Raw table	5 days	0.17	3.21	0.80	0.23	-0.05	-0.08
Raw table	20 days	0.68	6.68	2.68	0.75	0.01	-0.06
Chart	1 day	0.03	1.38	0.15	0.03	-0.05	-0.03
Chart	5 days	0.13	3.13	0.68	0.20	-0.11	-0.05
Chart	20 days	0.44	6.46	2.16	0.64	-0.15	-0.04
Summary	1 day	0.04	1.34	0.23	0.05	-0.04	-0.03
Summary	5 days	0.19	3.02	1.15	0.25	-0.09	-0.06
Summary	20 days	0.84	6.22	4.20	0.85	-0.07	-0.08
No reasoning							
Raw table	1 day	0.01	1.89	0.07	0.09	-0.08	-0.04
Raw table	5 days	0.10	4.20	0.39	0.39	-0.18	-0.03
Raw table	20 days	0.44	8.54	1.49	1.43	-0.33	0.01
Chart	1 day	0.00	1.32	0.04	0.06	-0.03	-0.01
Chart	5 days	0.07	2.96	0.29	0.23	-0.08	0.00
Chart	20 days	0.36	6.04	1.13	0.80	-0.11	0.05
Summary	1 day	0.04	1.34	0.21	0.08	-0.05	-0.03
Summary	5 days	0.23	3.01	1.06	0.38	-0.09	-0.03
Summary	20 days	1.03	6.20	4.19	1.50	-0.10	-0.04

Note: Returns and standard deviations are stated cumulative simple-return percentage points from the first forecast. The four final columns are regression slopes on standardized earlier and latest log returns, with both included in every regression. Each slope is measured in return percentage points per standard deviation of the respective past return.

Table F.3: Cue contrasts using the full displayed return

Contrast	1 day	5 days	20 days	Average	95% interval
Summary minus other formats: history, reasoning	2.73	4.99	7.43	5.05	[4.35, 5.75]
Summary minus other formats: history, no reasoning	3.98	5.06	6.90	5.31	[4.75, 5.88]
Reasoning minus no reasoning: history	2.53	4.70	6.50	4.58	[4.11, 5.05]
No reasoning minus reasoning: latest	1.41	1.30	1.81	1.51	[1.20, 1.81]
Summary acquisition: history reduction, reasoning	1.69	2.96	4.46	3.04	[2.58, 3.49]
Summary acquisition: history reduction, no reasoning	1.22	1.80	3.15	2.06	[1.83, 2.28]

Note: These regressions use the cumulative log return over all 19 displayed intervals and the latest one-day log return. The full-history regressor includes the latest return; this basis differs from the nonoverlapping earlier/latest specification in Table 3.3. Both regressors enter each regression and use their original global standard deviations. Entries are probability percentage points per standard deviation. Summary-minus-other-format contrasts subtract the equal average of raw-table and chart slopes. Acquisition contrasts average the two added representations. Intervals are simultaneous across the six rows, based on 4,999 paired state-stratified episode draws.

G Acquisition and financial decisions

Table G.1: Expected-payoff contrasts and uncertainty

Contrast	Estimate	95% interval	Numerical SE	First	Repeat
Reasoning minus no reasoning	67.4	[37.9, 96.9]	0.65	63.9	70.9
Reasoning minus affine adjustment	81.1	[50.9, 111.3]	0.69	79.0	83.2
Summary minus raw: no reasoning	15.6	[-19.6, 50.9]	0.87	26.7	4.6
Summary minus raw: reasoning	-15.1	[-45.8, 15.6]	0.56	-16.0	-14.2
Summary–raw reasoning interaction	-30.8	[-71.0, 9.5]	0.85	-42.7	-18.8
Reasoning gain: conflicting minus agreeing returns	113.8	[54.6, 172.9]	1.36	116.3	111.3
Summary acquisition: reasoning	20.6	[-4.5, 45.7]	0.45	—	—
Summary acquisition: no reasoning	-6.6	[-29.0, 15.8]	0.58	—	—
Summary acquisition: reasoning interaction	27.2	[-0.7, 55.1]	0.61	—	—
Summary acquisition minus affine adjustment	21.5	[-2.6, 45.6]	0.41	—	—

Note: All quantities are basis points of claim face value. Each horizon is evaluated separately before equal averaging. The first six contrasts also average the two independent initial forecasts. Summary acquisition averages the raw-table and chart additions to the initial summary. Intervals are simultaneous over the ten rows, using 4,999 paired state-stratified episode draws. The affine mappings are refitted within each draw using the retained training/test split. Conditional event probabilities are held fixed during resampling; the numerical standard error records their simulation error separately. Conflicting and agreeing groups exclude the five histories with an exact-zero earlier or latest return.

Table G.2: Realized acquisition payoffs at three position costs

Cost (bp)	Forecast	Gain (bp)	95% interval	Bootstrap SD
Reasoning				
0	Original	-2.55	[-90.12, 85.03]	28.55
0	Calibrated	40.56	[-62.62, 143.73]	33.64
0	Calibrated minus original	43.10	[-81.33, 167.54]	40.57
50	Original	16.57	[-63.35, 96.49]	26.06
50	Calibrated	49.27	[-34.76, 133.30]	27.40
50	Calibrated minus original	32.70	[-70.06, 135.45]	33.50
100	Original	18.38	[-54.47, 91.24]	23.75
100	Calibrated	41.99	[-33.53, 117.51]	24.62
100	Calibrated minus original	23.61	[-66.57, 113.79]	29.40
No reasoning				
0	Original	-67.33	[-172.37, 37.72]	34.25
0	Calibrated	37.37	[-79.81, 154.55]	38.20
0	Calibrated minus original	104.70	[-54.03, 263.42]	51.75
50	Original	-50.67	[-145.57, 44.23]	30.94
50	Calibrated	23.05	[-73.54, 119.63]	31.49
50	Calibrated minus original	73.72	[-64.47, 211.91]	45.06
100	Original	-34.15	[-120.47, 52.16]	28.14
100	Calibrated	32.84	[-53.31, 118.99]	28.09
100	Calibrated minus original	66.99	[-56.88, 190.86]	40.39

Note: Payoffs use the single realized binary outcome for each history. Gains subtract the initial payoff from the acquisition payoff and average the six ordered representation pairs and three horizons. Costs and gains are basis points of claim face value. Scalar calibration fits the realized outcome to the corresponding probability in five training/test splits. All required calibration maps are refitted in each of 4,999 paired state-stratified episode draws. Intervals are simultaneous across the 18 entries in both panels.

Table G.3: Acquisition, probability-loss components and financial positions

Measure	Unit	Reasoning: raw table	Reasoning: chart	Reasoning: summary	No reasoning: raw table	No reasoning: chart	No reasoning: summary
Original Brier gain	10^4	6.65	3.69	53.30	-11.00	-7.40	21.46
Squared mean-error component	10^4	-2.98	-4.34	-6.70	-8.47	-8.17	-11.27
Probability-variance component	10^4	13.22	7.73	72.85	0.60	-1.81	41.54
Outcome-covariance component	10^4	-3.59	0.29	-12.85	-3.13	2.59	-8.82
Calibrated Brier gain	10^4	0.56	1.76	2.38	-0.51	2.88	1.66
Expected-payoff gain	bp	-6.83	-17.83	20.60	-16.80	-39.29	-6.58
Positions changed	%	6.72	16.67	12.70	11.47	16.93	15.73
Payoff gain if position changes	bp	-101.66	-106.96	162.19	-146.52	-232.00	-41.85
Positive gain among changes	%	46.15	45.20	57.22	41.72	38.78	46.19

Note: Each numerical column identifies the receiver's reasoning condition and representation. Entries average the two added representations and the three horizons. Brier-loss reductions and their components are multiplied by 10^4. The mean-error, variance and covariance components sum to the original Brier gain; the covariance component is twice the change in covariance between forecast and outcome. Calibrated gains use separate scalar calibrations of initial and final probabilities. Position changes and expected payoffs use original forecasts, conditional event probabilities and a 100 bp position cost. Payoffs are basis points of claim face value. The last two rows condition on a change in the long, flat or short position.

Table G.4: Position changes and expected-payoff accounting

Sample	Position change	Frequency (%)	Gross gain (bp)	Cost effect (bp)	Net gain (bp)
Panel A. Reasoning minus no reasoning					
All histories	Long–short reversals	10.29	39.5	0.0	39.5
	Entries	10.32	14.3	-10.3	4.0
	Exits	9.04	11.4	9.0	20.4
	Unchanged	70.34	0.0	0.0	0.0
Earlier rise / latest fall	Long–short reversals	25.53	123.5	0.0	123.5
	Entries	7.13	12.4	-7.1	5.3
	Exits	18.28	50.0	18.3	68.2
	Unchanged	49.06	0.0	0.0	0.0
Panel B. Acquisition by summary receivers with reasoning					
All histories	Long–short reversals	3.18	11.7	0.0	11.7
	Entries	3.32	2.3	-3.3	-1.0
	Exits	6.20	3.7	6.2	9.9
	Unchanged	87.30	0.0	0.0	0.0
Earlier rise / latest fall	Long–short reversals	8.11	28.6	0.0	28.6
	Entries	7.04	-1.9	-7.0	-8.9
	Exits	11.70	0.2	11.7	11.9
	Unchanged	73.14	0.0	0.0	0.0

Note: Entries use first forecasts and a 100 bp cost. Contributions sum to the expected-payoff gain in the indicated comparison and sample. Frequencies are normalized within the sample. Panel A averages representations and horizons; Panel B averages the two representations added to a summary and all horizons. A positive cost effect is a reduction in position costs.

Table G.5: Reciprocal acquisition: probability, return and risk gaps

Pair	Horizon	Probability: initial	Probability: final	Return: initial	Return: final	SD: initial	SD: final
Reasoning							
Raw–chart	1 day	3.81	3.59	0.14	0.14	0.31	0.26
Raw–chart	5 days	6.17	5.47	0.55	0.48	0.67	0.58
Raw–chart	20 days	9.27	8.23	1.93	1.68	1.49	1.25
Raw–summary	1 day	4.55	3.77	0.16	0.14	0.34	0.20
Raw–summary	5 days	7.61	5.82	0.71	0.53	0.75	0.47
Raw–summary	20 days	11.53	8.76	2.98	1.96	1.65	1.12
Chart–summary	1 day	4.92	3.42	0.17	0.12	0.30	0.15
Chart–summary	5 days	8.42	5.63	0.75	0.49	0.66	0.34
Chart–summary	20 days	13.01	8.71	3.13	1.67	1.37	0.77
No reasoning							
Raw–chart	1 day	2.78	2.19	0.10	0.09	0.71	0.57
Raw–chart	5 days	3.61	3.03	0.42	0.37	1.51	1.23
Raw–chart	20 days	5.79	4.94	1.59	1.39	3.01	2.53
Raw–summary	1 day	5.12	3.71	0.19	0.13	0.69	0.40
Raw–summary	5 days	6.93	4.87	0.90	0.60	1.47	0.87
Raw–summary	20 days	10.20	6.54	3.63	2.18	2.87	1.77
Chart–summary	1 day	6.02	3.80	0.22	0.13	0.42	0.17
Chart–summary	5 days	7.58	4.76	1.01	0.62	0.90	0.39
Chart–summary	20 days	10.92	6.52	4.05	2.36	1.76	0.88

Note: Each entry is a root mean squared gap between reciprocal representation routes. Initial gaps use the first forecasts. Final gaps compare forecasts after each route acquires the other representation. Probability columns are probability percentage points; return and SD columns are stated simple-return percentage points. Final requests contain the same two representation sources, each route's own initial forecast, and its presentation order.

Table G.6: Past returns and reciprocal forecast gaps

Pair	Horizon	Probability earlier: initial	Probability earlier: final	Return earlier: initial	Return earlier: final	Probability latest: initial	Probability latest: final
Reasoning							
Raw–chart	1 day	0.63	0.53	0.02	0.02	0.21	0.24
Raw–chart	5 days	1.24	1.07	0.12	0.11	0.50	0.51
Raw–chart	20 days	1.95	1.80	0.53	0.47	0.72	0.85
Raw–summary	1 day	-2.32	-1.00	-0.07	-0.02	-0.37	-0.47
Raw–summary	5 days	-4.22	-1.96	-0.35	-0.12	-0.38	-0.84
Raw–summary	20 days	-6.22	-3.00	-1.51	-0.47	-0.57	-0.95
Chart–summary	1 day	-2.95	-0.94	-0.09	-0.02	-0.59	-0.56
Chart–summary	5 days	-5.45	-2.19	-0.47	-0.14	-0.88	-1.03
Chart–summary	20 days	-8.17	-3.77	-2.04	-0.68	-1.29	-1.49
No reasoning							
Raw–chart	1 day	0.51	0.48	0.03	0.03	0.62	0.33
Raw–chart	5 days	0.33	0.38	0.09	0.09	0.96	0.76
Raw–chart	20 days	0.43	0.61	0.36	0.35	1.49	1.36
Raw–summary	1 day	-3.60	-2.30	-0.14	-0.08	0.11	-0.12
Raw–summary	5 days	-4.73	-2.85	-0.67	-0.37	0.18	-0.14
Raw–summary	20 days	-6.46	-3.35	-2.70	-1.35	0.33	0.01
Chart–summary	1 day	-4.10	-2.42	-0.17	-0.09	-0.50	-0.19
Chart–summary	5 days	-5.07	-2.88	-0.76	-0.42	-0.77	-0.54
Chart–summary	20 days	-6.89	-3.69	-3.06	-1.68	-1.16	-0.70

Note: Entries are regression coefficients of the signed representation gap on standardized earlier and latest log returns, including an intercept and both regressors. A pair is oriented as its first representation minus its second. Initial gaps use the first forecasts; final gaps use reciprocal acquisition forecasts. Probability coefficients are probability percentage points per standard deviation of the respective past return; return coefficients are stated simple-return percentage points per standard deviation.

H Consultation and the full forecast report

Table H.1 records conditional expected losses and payoff gains for all nine ordered consultation pairs. Table H.5 records simultaneous payoff intervals. Table H.2 reports original continuous-loss accounting, and Table H.3 reports fitted return-loss comparisons. Table H.4 records directional and cumulative log-return predictions based on one probability, all three probabilities and all nine reported fields. Conditional Brier loss is higher for the nine-field mapping than for the three-probability mapping at each horizon under both reasoning conditions. The eight simultaneous calibrated probability-loss intervals in Table E.3 include zero.

The original continuous loss is squared error in cumulative simple-return percentage points divided by the squared fixed horizon-specific reference scale. The reference scales and public-prior returns are fixed simulation-derived constants. Under the untruncated lognormal variance law, conditional Gaussian integration yields $\mathbb{E}[e^{ar_t} \mid v_t, \text{other states}] = \exp(a\,\text{conditional mean} + a^2 v_t/2)$ for $a > 0$. A nondegenerate lognormal v_t has $\mathbb{E}[e^{cv_t}] = \infty$ for $c > 0$. Conditioning on the final day's preceding history establishes that $R_h = e^{\ell_h} - 1$ has infinite first and second population moments at each evaluated horizon. The original simple-return losses and their resampling intervals are statistics of the retained finite sample. The directional outcome is bounded. The cumulative log-return predictions in Table H.4 use transformed realized return labels and report root mean squared errors in log-return percentage points. Stated expected returns and stated standard deviations remain in their elicited simple-return units.

Full-report return mappings use nine fields per report: the probability, expected return and stated standard deviation at each horizon. The continuous-loss mappings in Table H.3 select a ridge penalty from $\{0.001, 0.01, 0.1, 1, 10\}$ using four inner training folds. The report-content mappings in Table H.4 use the fixed penalty one. Predictor centering and scaling use training observations only. All reported prediction losses use held-out episodes.

Table H.1: Consultation by receiver and peer representation

Receiver	Peer	Initial loss	Pool loss	Final loss	Pool gain (bp)	Final gain (bp)
Reasoning						
Raw table	Raw table	2,546.43	2,542.62	2,545.39	10.32	4.26
Chart	Chart	2,518.57	2,509.59	2,508.84	12.28	9.75
Summary	Summary	2,621.94	2,617.76	2,623.46	-0.55	-0.97
Raw table	Chart	2,546.43	2,520.64	2,520.75	19.66	10.99
Raw table	Summary	2,546.43	2,570.18	2,569.07	0.79	-4.53
Chart	Raw table	2,518.57	2,524.98	2,530.24	-5.38	-8.52
Chart	Summary	2,518.57	2,550.20	2,547.90	-11.49	-9.01
Summary	Raw table	2,621.94	2,571.57	2,586.13	10.95	11.80
Summary	Chart	2,621.94	2,547.88	2,556.75	12.90	10.04
No reasoning						
Raw table	Raw table	2,512.82	2,510.75	2,509.54	4.62	12.11
Chart	Chart	2,500.79	2,499.81	2,498.94	-6.03	-6.55
Summary	Summary	2,551.84	2,549.67	2,548.76	-12.04	-12.58
Raw table	Chart	2,512.82	2,502.48	2,501.71	-2.37	0.63
Raw table	Summary	2,512.82	2,518.30	2,518.61	11.47	12.32
Chart	Raw table	2,500.79	2,502.11	2,503.18	-13.62	-7.29
Chart	Summary	2,500.79	2,509.60	2,510.38	6.82	5.18
Summary	Raw table	2,551.84	2,517.26	2,518.89	-9.42	-9.33
Summary	Chart	2,551.84	2,509.12	2,511.11	1.47	-0.70

Note: Losses are conditional expected Brier losses multiplied by 10^4. Payoff gains compare the equal pool or final consultation forecast with the matched initial forecast, in basis points of claim face value. Each ordered pair averages the three horizons. Peer forecasts are from the independent repeat. A 100 bp cost is applied to each nonzero position.

Table H.2: Consultation diversity in continuous forecast losses

Reasoning	Horizon	Total	Averaging	Revision
Reasoning	1 day	0.18	0.14	0.04
	5 days	0.53	0.57	-0.04
	20 days	1.86	1.83	0.03
	Average	0.86	0.85	0.01
No reasoning	1 day	0.31	0.33	-0.02
	5 days	1.27	1.27	0.00
	20 days	4.64	3.84	0.79
	Average	2.07	1.82	0.26

Note: Entries subtract same-format from different-format consultation gains in normalized squared simple-return loss, multiplied by 100. Total gain equals averaging plus subsequent revision. The fixed horizon-specific return scales are 1.4161, 3.3248 and 7.6617 percentage points. Losses are finite-sample quantities relative to these reference scales.

Table H.3: Fitted continuous forecast losses

Reasoning	Forecast mapping	Same format	Different format	Same minus different
Reasoning	Calibrated equal pool	97.89	97.81	0.08
Reasoning	Two return forecasts	98.12	98.03	0.09
Reasoning	Two full reports	98.91	98.62	0.29
Reasoning	Final return forecast	97.96	97.85	0.11
Reasoning	Final full report	98.42	98.22	0.20
No reasoning	Calibrated equal pool	97.73	97.59	0.14
No reasoning	Two return forecasts	97.71	97.76	-0.05
No reasoning	Two full reports	98.21	98.01	0.20
No reasoning	Final return forecast	97.73	97.60	0.13
No reasoning	Final full report	98.20	97.87	0.33

Note: Entries are normalized squared simple-return losses or differences multiplied by 100, averaged equally across horizons. Mappings use five-fold held-out evaluation. Scalar mappings use an intercept and one or two expected-return reports; full-report mappings use all nine fields from each available report and ridge regression with a penalty selected in four inner folds. The fixed return scales are those in Table H.2.

Table H.4: Forecast content in probabilities, expected returns and risk

Reasoning	Horizon	Report fields	Conditional Brier loss	Expected payoff (bp)	Log-return RMSE (pp)
Reasoning	1 day	One probability	2,484.60	85.70	1.40
		Three probabilities	2,483.35	84.44	1.40
		Full report	2,495.09	33.81	1.41
	5 days	One probability	2,481.41	127.32	3.32
		Three probabilities	2,483.05	107.22	3.31
		Full report	2,494.92	81.84	3.33
	20 days	One probability	2,444.75	162.20	7.68
		Three probabilities	2,445.63	159.33	7.66
		Full report	2,460.71	94.10	7.68
No reasoning	1 day	One probability	2,493.04	-1.20	1.40
		Three probabilities	2,493.23	33.94	1.40
		Full report	2,501.08	28.78	1.40
	5 days	One probability	2,485.75	92.53	3.32
		Three probabilities	2,488.43	73.70	3.32
		Full report	2,498.71	48.00	3.33
	20 days	One probability	2,447.94	129.84	7.66
		Three probabilities	2,449.73	135.66	7.67
		Full report	2,462.33	100.27	7.71

Note: One probability uses the target horizon; three probabilities use all reported horizons; the full report adds the three expected returns and three stated standard deviations. Each target is fitted separately using an intercept, training-standardized predictors and ridge regression with fixed penalty one, with five-fold held-out evaluation. Brier loss is multiplied by 10^4; payoff is basis points of claim face value at a 100 bp position cost. Cumulative log-return RMSE is in percentage points. The corresponding training-mean RMSEs at one, five and twenty days are 1.40, 3.35, 7.75 with reasoning and 1.40, 3.35, 7.75 without reasoning. Entries average the three representations and use first reports.

Table H.5: Consultation payoff gains and uncertainty

Reasoning	Comparison	Estimate (bp)	95% interval
Reasoning	Final minus initial	1.8	[-10.0, 13.6]
	Pool minus initial	4.6	[-7.3, 16.4]
	Final minus pool	-2.8	[-6.6, 1.0]
	Different minus same	-2.6	[-10.8, 5.6]
No reasoning	Final minus initial	0.1	[-11.5, 11.8]
	Pool minus initial	-0.9	[-12.5, 10.6]
	Final minus pool	1.1	[-1.5, 3.7]
	Different minus same	2.5	[-7.0, 12.0]

Note: Payoff gains are basis points of claim face value at a 100 bp position cost, averaged across horizons. The first three comparisons use different-format peers. Different-minus-same subtracts the corresponding final-minus-initial gains. Intervals are simultaneous across all eight entries. Episode resampling holds the stored conditional event probabilities fixed.