Completions, Not Convictions: Can LLM Probabilities Proxy for Human Beliefs?

Stefan Nagel
University of Chicago, NBER, CEPR, and CESIfo
stefan.nagel@chicagobooth.edu

Claire Tseng
University of Chicago
clairetseng@uchicago.edu

Dacheng Xiu
University of Chicago and NBER
dacheng.xiu@chicagobooth.edu

This Version: August 2026

Acknowledgments. We thank seminar and conference participants for helpful comments. We gratefully acknowledge computing support from the University of Chicago's Research Computing Center and financial support from the SCOR Foundation for Science.

Completions, Not Convictions:

Can LLM Probabilities Proxy for Human Beliefs?

This Version: August 2026

Abstract

An ideal *homo silicus* would let economists elicit beliefs at scale. In prediction markets and analyst earnings, however, LLM probabilities deviate systematically from human consensus. As prompt continuations, they respond to tone, anchors, label order, training-corpus exposure, and outcome direction. Richer information, demographic aggregation, reasoning instructions, and model scale shift the gap without reliably closing it. Substituting LLM probabilities for analyst beliefs changes their relation with announcement returns and reverses forecast-error-on-revision conclusions: analysts underreact, while the LLM appears to overreact. LLM probabilities therefore require model-, prompt-, setting-, and use-specific validation and are not human beliefs by construction in economic research.

1 Introduction

Understanding how human agents form beliefs is fundamental to economics and social science, but measuring those beliefs at scale remains difficult. Surveys are expensive, experimental subjects are difficult to recruit, and the conditions that let researchers cleanly identify beliefs are often unavailable in natural settings. Large language models (LLMs) offer a potential solution: trained on large corpora of human-generated text, they may encode enough about human reasoning, preferences, and judgment to serve as synthetic research participants. If their responses approximate what humans would produce, LLMs could substitute for human respondents at scales and speeds beyond what field methods allow.

The object being substituted for can be an individual respondent, a demographic group, a population aggregate, or a market participant. Current practice endows the model with information meant to help it respond as the relevant person or group would; these inputs include *who* the forecaster is, set by instruction, and *what* the forecaster sees, set by supplied information. This approach, in which the model serves as a computational stand-in for human responses, is known as *homo silicus* (Horton et al., 2023). The natural question is when a probability elicited from the model is a valid substitute for the corresponding human belief.

The premise has a concrete basis in how these models are trained, although training does not explicitly target the reproduction of human judgment. LLMs are pre-trained on human text and commonly post-trained to produce responses that human raters prefer (Ouyang et al., 2022). Aligning a model to human *preferences*, however, need not align its elicited *probabilities* to human *beliefs*. Such tuning optimizes for answers that raters score highly

rather than for calibrated event probabilities, and calibration can degrade after alignment (OpenAI, 2023). The difference matters: a number that validly measures belief can replace a human's at scale, while one that partly reflects how it was elicited can corrupt analyses built on the substitution.

Forecasting settings can separate a valid belief measure from an elicitation-driven number because they supply the external benchmark the homo silicus premise requires: the human probability and the realized outcome. We study two such settings, prediction markets and analyst earnings forecasts. The two settings are complementary along several dimensions. First, they instantiate two ways of endowing a homo silicus in our design: behavioral cueing through investor personas in prediction markets and informational cueing through recent earnings per share (EPS) in analyst earnings, with firm disclosures used later to test richer context. Second, the human benchmarks are fundamentally different aggregation mechanisms: a market price formed by anonymous traders with money at stake in prediction markets and a professional consensus of named analysts with career incentives in analyst earnings. Comparing the LLM's forecasts against both sheds light on whether its biases reflect the nature of the benchmark or properties of the model. Third, they differ in horizon and coverage breadth: prediction-market contracts resolve in days to weeks and span topics from macroeconomic releases to sports to weather, while analyst earnings forecasts extend up to a year and cover a dense cross-section of firms and sectors. Together they let us examine whether the gap and its components vary with event characteristics (topic, sector, coverage intensity, horizon) rather than being an artifact of one institutional environment.

Within each setting, we ask LLaMA 3.3-70B to assign probabilities using inputs that mirror what the human forecaster would have; we refer to this specific model simply as

"the LLM". Both humans and the LLM assign probabilities to events that later resolve, so the human probability is the natural target for the belief-proxy claim while the realized outcome lets us judge both forecasters objectively. We compare them using each forecaster's outcome-aligned probability, the probability assigned to the side of the true outcome. In prediction markets, where we simulate a crowd with different demographic personas, the LLM's outcome-aligned probability is about 23 percentage points (pp) below the human benchmark.

Enriching the prompt with what the human forecaster sees does not close the gap. In analyst earnings the prompt supplies the firm's recent EPS, and further adding firm disclosures does not close it further: neutral Management's Discussion and Analysis (MD&A) excerpts leave the gap nearly unchanged, while Risk Factors and Forward-Looking Statements widen it by 13 and 12 pp relative to that EPS baseline. In prediction markets we supply a numerical probability directly: the previous day's market price closes about half the gap. The forecast follows the supplied number even when it is a random value or the complement of the true price, so the model is anchoring on the figure rather than reasoning from it.

Part of the gap surely reflects information the human has and the prompt lacks, and richer context is a natural remedy. The false anchors suggest that a gap closing under richer context would be as consistent with the model following the prompt as with the model reproducing the belief. Setting that aside, attributing the gap entirely to missing information leaves validity contingent on information a researcher cannot observe: in prediction markets, the private signals of thousands of traders; in analyst earnings, the expertise an analyst accumulates over a career. A prompt cannot plausibly carry this information, so the claim that fuller information would close the gap can be checked only against human probabilities,

which are the data the substitution is meant to replace. The gap does not behave as a shortfall of information in any case: some of the events we study resolved before the model was trained, so their outcomes could appear in the text it learned from, an advantage no forecaster had at the time. Even on those events the model assigns 18.9 pp less probability to the realized outcome than the human benchmark in prediction markets and 14.2 pp less in analyst earnings, and the gap is wider still on events whose outcomes it could not have seen.

The structure also persists when we vary the elicitation rather than the information. In our design, relabeled or reordered response options pose the identical question, so any movement in the probability reflects the elicitation form rather than a change in belief. The LLM's stated probabilities move materially with both: reordering labels flips the directional call on up to 11% of prediction-market events and shifts analyst forecasts systematically toward one label. Reasoning instructions close at most 6% of the remaining gap, and larger or different model families change which component dominates without eliminating the structure.

Underlying these patterns is what the elicited probability is in our design: a continuation of the prompt, not a directly observed belief the model stores and reports. Within this object the model tracks only part of the human probability variation and otherwise falls back toward model-side components. An LLM probability must therefore be evaluated as the output of a specific model, prompt, setting, and downstream use.

Our findings speak to several literatures. Researchers increasingly deploy LLMs as stand-ins for human respondents: recent economic and financial applications use demographically seeded or calibrated LLM agents to study investment preferences and investor reactions to firm news (Fedyk et al., 2024; Bhagwat et al., 2026). Several of the phenomena we document

have precedents. The AI evaluation literature documents that stated answers shift with the order and labeling of response options (Zheng et al., 2024) and that benchmark scores are inflated by train–test contamination, though it treats these as isolated engineering defects rather than as evidence about whether an elicited probability is a valid measure of human belief. The survey-methodology literature has long shown that question form moves human responses (Tversky and Kahneman, 1981; Schuman and Presser, 1981; Manski, 2004); our benchmarks, a market price and an analyst consensus, are formed without any such instrument, so the form effects we document belong to the proxy rather than to human framing. We bring these phenomena together against economic ground truth. The two settings we use are also of independent interest: prediction markets have become an increasingly prominent mechanism for aggregating expectations about political, macroeconomic, and financial outcomes (Ng et al., 2026; Diercks et al., 2026), while analyst earnings forecasts are a central setting for studying how professionals form beliefs and process firm-level information (Bastianello et al., 2024; Ke, 2024; Bastianello, 2025).

Our exercise also sits in a broader effort to use imperfect machine-generated outputs as data without letting their errors corrupt the conclusions drawn from them: ML classifiers that construct binary labels can bias the downstream regressions that rely on them (Goldsmith-Pinkham and Shue, 2023; Bursztyn et al., 2024; Dupas et al., 2026; Hansen et al., 2023; Christensen et al., 2026). A related study extends this concern to measures generated by the LLM itself: Grieser et al. (2026) study text measures built from LLM summaries and report non-classical measurement error, including omitted source content and injected vocabulary that reflects what was salient in the model's training period, along with diagnostics for the resulting bias in downstream regressions. A separate line of work elicits continuous beliefs

from LLMs rather than discrete labels, generating economic expectations from news text and benchmarking them against human survey measures (Bybee, 2025). Our continuous object differs in where it comes from: rather than a number the model states in words, it is the model's continuation probability, the probability it places on the answer token that continues the prompt, read directly from its next-token distribution and used as a proxy for a human belief.

Our constructive goal, to diagnose the bias in a machine-generated signal and correct it with a limited amount of trusted data, is shared in spirit with prediction-powered inference (Angelopoulos et al., 2023), which uses a small validation sample of human labels to restore valid inference on parameters estimated from machine predictions. A validation sample of human benchmarks plays that role here: it pins down the exposure, directional, and form components of the gap, the error model through which a new elicitation is mapped into an estimate of the human belief. Training-corpus exposure connects to a growing literature on lookahead bias in LLM forecasting (Sarkar and Vafa, 2024; Lopez-Lira et al., 2025; Levy, 2026): existing approaches address it by retraining the model on chronologically structured data (He et al., 2025; Sarkar, 2024) or anonymizing event-identifying information (Glasserman and Lin, 2023), whereas our diagnostics require no retraining or anonymization. Concurrent work by Gao et al. (2025) likewise modifies neither model nor data: they detect lookahead bias by measuring the model's recall of realized outcomes and showing that its forecast accuracy rises with this recall measure only before the training cutoff. Our exposure diagnostic shares the cutoff logic and embeds it in the decomposition of the gap, where exposure is estimated jointly with the directional and form components.

A final remark concerns scope. For demonstration and reproducibility, we use a sin-

gle fixed, open-source model, so that every elicitation and result can be replicated exactly. In cross-model results for LLaMA 3.1-8B and Qwen 2.5, training-corpus exposure and directional asymmetry recur, but their signs, magnitudes, and precision vary by model and setting: no pattern is universal enough to carry over from one system to the next. Rather than resolving these concerns, agentic systems can propagate the biases we document. They retrieve text, take numbers from tools, and reason over many steps, and we find bias in each of these three channels. The practical implication is that diagnosis must precede elicitation: the components have to be estimated for the model, elicitation, and setting actually deployed, whether an open model like ours or a more advanced proprietary system.

The remainder of the paper is organized as follows. Section 2 introduces the elicited probability, the two settings, and the human benchmarks. Section 3 documents the LLM–human gap and the limits of persona aggregation and added context. Section 4 decomposes the gap into training-corpus exposure and a directional prior, and shows how these components vary with model scale and family. Section 5 then holds the event fixed and measures how the elicited probability responds to the prompt itself: firm disclosures, supplied probabilities and false anchors, label order and wording, and reasoning instructions. Section 6 tests whether the raw probability and linear corrected scores reproduce human-forecast inferences in announcement-return and belief-rigidity regressions. Section 7 concludes. The appendix provides prompt details, identification and robustness evidence, and additional heterogeneity and cross-model results.

2 Elicited Probabilities, Settings, and Benchmarks

We construct a panel of probabilistic forecasts generated by a fixed LLM across two settings, prediction markets and analyst earnings forecasts, both of which have outcomes that resolve on known dates. We compare each to a human benchmark, holding the elicitation fixed throughout so that forecasts differ only in the information the prompt carries. The fixed model, LLaMA 3.3-70B, is instruction-tuned: beyond pre-training, it is post-trained through supervised fine-tuning and preference optimization, methods designed to produce responses that human raters prefer (Ouyang et al., 2022; Grattafiori et al., 2024).

2.1 Defining the Elicited Probability

The elicited probability is a deterministic function of the entire prompt: the model defines a distribution over the next token, and a *single elicitation* yields this probability as the softmax of that distribution restricted to the two candidate answer tokens.[1] With the prompt written as a token sequence x and the two answer tokens a_{fav} and a_{unf}, this probability is

$$p(x) \;=\; \frac{\exp\{z_{a_{\text{fav}}}(x)\}}{\exp\{z_{a_{\text{fav}}}(x)\} + \exp\{z_{a_{\text{unf}}}(x)\}}, \tag{1}$$

where $z_a(x)$ is the model's logit for token a conditional on x. Here a_{fav} and a_{unf} are the tokens for the favorable and unfavorable answers, which correspond to Yes/Increase and

[1] Candidate responses are the leading tokens of the two outcome labels: "Yes" and "No" for prediction markets, "Increase" and "Decrease" for analyst earnings. Extracting probabilities directly from token-level log-likelihoods rather than from free-text generation produces a well-defined conditional probability over these two labels for every observation; there are no refusals or non-responses to handle. Chen et al. (2026) likewise extract token-level conditional ("inner") probabilities from an LLM and show that their relative ranking carries economically meaningful predictive information in financial news classification and portfolio formation, supporting the use of these probabilities as economically informative model outputs.

No/Decrease, respectively.[2] Whether this probability measures a belief about the event is an empirical question, and the answer depends on three properties of the model that distinguish it from such a belief.

First, training fits the logits on a corpus collected up to a cutoff date: when an event resolved before the cutoff and generated dense post-resolution text, that text can raise $z_a(x)$ on the realized side, so the elicitation may reflect training-corpus exposure rather than out-of-sample forecasting, and because exposure operates only on outcomes with distinctive coverage, it can be direction-asymmetric. Second, text appended to x enters the logits through its tokens' learned associations as well as through its event-relevant content; this is the tone-dependence. Third, $p(x)$ depends on the wording and ordering of the response labels: two prompts that pose the same question with the labels reordered or reworded are different sequences x, so they can return different $p(x)$ even though a human reads them as identical. This prompt sensitivity separates into a positional channel (the logits are not invariant to token order) and a connotation channel (tokens such as "Yes" and "Increase" carry favorable mass accumulated in training); the population average of the connotation channel is a structural tilt in $z_{a_{\text{fav}}} - z_{a_{\text{unf}}}$, present before any event-specific information enters x.

All three follow from one fact: the elicited probability is a functional of the prompt's token sequence, so it responds to form as well as content, whereas the belief it is meant to measure is defined over the event. They are different objects, and the rest of the paper measures the gap between them.

[2] *Favorable* is a setting-neutral label: exact in analyst earnings, where it is an EPS increase, but a shorthand in prediction markets, where a YES is only the literal "Yes" answer and can itself be bad news.

2.2 Prediction Markets and Analyst Earnings

2.2.1 Prediction Markets

We collect the prediction-market data from Polymarket and Kalshi, covering markets with resolution dates from July 2021 (Kalshi) and December 2022 (Polymarket) through April 2025. Each market is a binary contract with an explicit resolution criterion and a fixed resolution date, which we take to be the scheduled close date, since that is when the underlying event resolves and its outcome becomes publicly known.[3]

The human benchmark is the daily market price, which equals the market's aggregate probability of a YES resolution.[4] Throughout, a market's price at a given date is its most recent trade as of that date. Polymarket trades continuously, so this is effectively the contemporaneous price; Kalshi contracts trade intermittently, so a Kalshi price is the last trade at or before the date. The estimation sample keeps contracts that resolved to a definite YES or NO, each observed over the 30 days before resolution, since earlier periods have thin trading and unreliable price discovery. On Kalshi, where a single event can bundle several contracts, we keep only single-leg events and exclude multi-leg families (mutually exclusive contracts that share a resolution, such as an election's candidate slate), whose legs are mechanically correlated and dominated by low base-rate NO outcomes, so they are not independent forecasts. We also drop short-duration contracts (under seven days), which tend to

[3] Although Polymarket dates from 2019, contract data was only meaningfully migrated onto its current system in 2022. We use the close date (the `end_date_iso` field for Polymarket and `close_time` for Kalshi) rather than the platform's settlement timestamp, which records the contract payout, can lag by hours or days during processing, and does not reflect when the outcome was knowable.

[4] Strictly, the price of a binary contract is a risk-neutral probability. We treat it as a proxy for the physical probability because the contracts are short-dated and most outcomes are largely uncorrelated with aggregate wealth, so the risk adjustment is second order. Consistent with this, the price is nearly perfectly calibrated against realized outcomes in our sample (Section 2.3).

be niche, thinly traded, and noisily priced; Kalshi's recurring series that repeat near-identical questions, such as daily weather and tweet- or mention-count markets; and Polymarket contracts whose price feed goes stale, last updating more than a year before the scheduled close. These restrictions, with the requirement of a valid forecast at each evaluation date, yield the estimation panel used throughout: 8,131 markets (7,208 Polymarket and 923 Kalshi), 158,814 market-date observations, and 815 markets resolving before the training cutoff. The panel is unbalanced: the median market is observed on 18 dates, and fewer than a third are observed on every date of the window. Most of the missing market-dates come from coverage that stops early rather than from holes inside a market's span: for 2,364 markets the price series ends before the scheduled close, a median of eight days early, with no later price in the source data. Markets that stop early are thinner, end nearer a certain price, and resolve YES more often, so coverage thins toward resolution in a way that is not random. The specifications below include days-to-resolution effects, which absorb this composition. For the heterogeneity analyses, we sort each prediction market into one of ten Kalshi topic categories (Appendix A.7); we use these categories only descriptively, and they do not enter prompting or exposure assignment.

2.2.2 Analyst Earnings and Firm Disclosures

The forecasting target is the year-over-year direction of EPS. For each firm-year i (a firm–fiscal-year pair), we ask the LLM and analysts whether reported EPS for that fiscal year will be higher or lower than reported EPS for the prior fiscal year; the realized outcome is the actual sign of the change, observed once per firm-year on the annual earnings announcement date. The analyst earnings sample covers S&P 500 firms over 2020–2025, using LSEG IBES

(Institutional Brokers' Estimate System) Detailed Estimates for analyst-level forecasts and Compustat for firm fundamentals, merged at the firm-year level. Index membership changes over time, so we keep a firm-year only if the firm was in the S&P 500 for the entire fiscal year, and we drop firm-years affected by a merger, acquisition, bankruptcy, or major restructuring.

The raw human benchmark $\tilde{p}_{it}^{H}$ is the fraction of analysts forecasting an increase:

$$\tilde{p}_{it}^{H} = \tfrac{1}{|\mathcal{F}_{i,t}|} \sum_{j \in \mathcal{F}_{i,t}} \mathbf{1}\{\hat{e}_{j,i,t} > E_i^{\text{prior}}\},$$

where t indexes the evaluation date, $\mathcal{F}_{i,t}$ is the set of analysts with an active forecast for firm-year i as of date t, $\hat{e}_{j,i,t}$ is analyst j's point forecast of firm-year i's EPS as of date t, and E_i^{prior} is firm i's realized EPS for the prior fiscal year. This consensus updates as new information arrives, converging toward the realized outcome as analysts revise initially optimistic forecasts toward it (Richardson et al., 2004; Easterwood and Nutt, 1999; Matsumoto, 2002), and at its final update assigns more than one-half probability to the realized EPS direction for 95.0% of firm-years. We date a firm-year by that final update and measure event time against it. Because IBES records when an estimate is issued, this *resolution date* precedes the fiscal-period end, by a median of fifteen days and never after it.

We keep firm-years with a median of at least five active analysts and a forecast horizon of at least thirty days, retain evaluation dates within one year of that resolution date, and require a Compustat record whose Global Industry Classification Standard (GICS) sector is available and whose identity matches the IBES ticker.[5] Because Compustat coverage

[5]This requirement excludes 47 firms whose IBES ticker returns no Compustat record, and hence no sector code, together with four whose ticker resolves to an unrelated Compustat company; where one company carries two ticker records in a fiscal year we keep the better-covered one (thirteen firm-years). Standard errors are clustered on the Compustat company identifier rather than the ticker.

ends when a firm ceases to trade, this requirement conditions the panel on surviving firms. Firm-level controls from Compustat capture size (log assets), profitability (return on assets), valuation (book-to-market), and cash holdings, each taken from the most recent quarterly report released on or before the evaluation date. The resulting panel comprises 2,649 firm-years drawn from 579 distinct companies (592 tickers; 132,957 firm-year–date observations), each observed at an average of 50.2 analyst-update dates over the forecast horizon.

We also use Form 10-K (annual report) and 10-Q (quarterly report) filings from SEC EDGAR, 2020–2025, for the same S&P 500 firms. We extract two narrative sections that firms must include in these filings. *Management's Discussion and Analysis* is management's own account of the period's financial condition and results of operations, covering performance, liquidity, and the trends behind them; *Risk Factors* is the mandated enumeration of the material risks and uncertainties that could adversely affect the business, spanning competitive, regulatory, macroeconomic, and operational threats. After removing tables, boilerplate, and HTML formatting, we construct four disclosure conditions, three drawn from the MD&A and one from Risk Factors. From the MD&A we draw two variants (a *Sampled* excerpt taken at random from across the section and a *Truncated* excerpt from its start), which together test sensitivity to which portion of the same content is supplied; we isolate the MD&A's *Forward-Looking Statements*, its projections and guidance about future results, as a third condition; and we take *Risk Factors* as the fourth.

2.3 Elicitation, Baselines, and Forecast Calibration

Each question is a structured prompt asking for the likelihood of the favorable resolution, and yields the raw probability $\tilde{p}_{it}^{\mathrm{LLM}}$ of Equation 1. The bare prompt, containing only the

evaluation date and the forecasting question, is the *minimal elicitation*, which the informed baselines defined next and the treatments in later sections all extend. For prediction markets we elicit at multiple dates while the market is active, mirroring how participants repeatedly update beliefs about a fixed event; for analyst earnings, we query each firm-year on each analyst-update date. We prompt each date independently and the realized outcome is never provided to the LLM.[6]

In prediction markets the human probability represents a crowd, so for the baseline we construct a *demographic* persona crowd, following Bhagwat et al. (2026). Each persona is a one-sentence demographic descriptor prepended to the prompt. We draw five traits (age, gender, race, income, and education) from FINRA's National Financial Capability Study and cross them with political orientation, which also moves beliefs in Horton et al. (2023). To these we add financial risk tolerance, a recognized driver of investor heterogeneity (Fedyk et al., 2024). For each market-date we draw a subsample of $K = 54$ personas balanced across the seven traits, rotating through the full pool of 1,296.[7] The crowd averages the single elicitations over each date's block,

$$\tilde{p}_{it}^{\text{LLM}} = \frac{1}{K} \sum_{k=1}^{K} p(x_{k,it}),$$

where $x_{k,it}$ prepends persona k to the market-date it prompt and $p(\cdot)$ is the single elicitation of Equation 1. In analyst earnings, following the practice of endowing the model with recent,

[6] Verbatim prompts for the baseline and for each later deviation (label-ordering variants, reasoning conditions, and context conditions) are reproduced in Appendix A.

[7] The 144 demographic cells, crossed with three political orientations and three risk-tolerance levels, give the 1,296 personas; 24 divides the cells evenly, so each block of $K = 1{,}296/24 = 54$ personas is balanced across the seven traits. Block assignment is a deterministic hash of the market identifier, hence independent of the outcome and exposure, so the daily subsampling adds noise to the crowd mean without biasing the gap. Appendix A.3 gives the block construction and rotation.

structural data as Hansen et al. (2026) do, we provide the LLM with a single elicitation whose prompt contains the firm's recent quarterly and yearly EPS,

$$\tilde{p}_{it}^{\mathrm{LLM}} = p\big(x_{it}^{\mathrm{EPS}}\big)\,.$$

These two prompts are the paper's *informed baselines*: each is a structured stand-in for the human forecaster, following the homo-silicus literature.

Comparing them to the human benchmark is informative only if that benchmark is itself a reliable forecaster. The forecast-calibration regression of Mincer and Zarnowitz (1969) regresses the realized-outcome indicator $g_i \in \{0, 1\}$, equal to one when the event resolves in the *favorable* direction and zero otherwise, on the raw stated probability $\tilde{p}_{it}^{f}$ for each forecaster $f \in \{\mathrm{LLM}, H\}$:

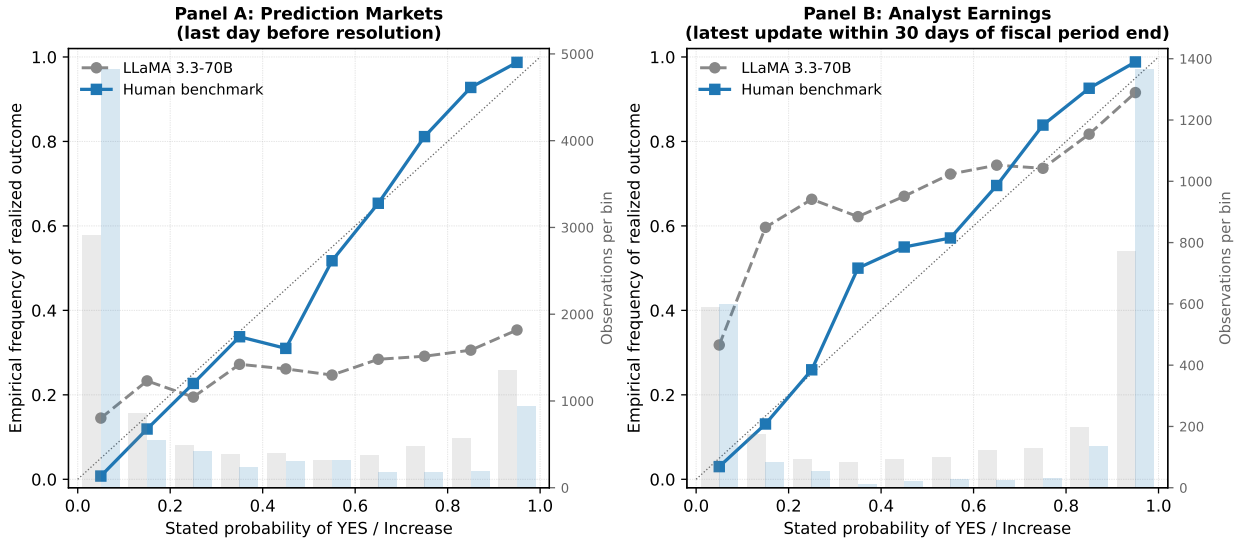
$$g_i = \alpha_f + \beta_f\, \tilde{p}_{it}^{f} + \varepsilon_i,$$

evaluated at each forecast's most informed date: the last pre-resolution trading day for prediction markets and the latest analyst update within thirty days of the fiscal period end. On the firm-years the window excludes, the analyst's Mincer–Zarnowitz slope falls from 1.00 to 0.92: their consensus was last updated a median of fifty days before the period end rather than twelve. The LLM's slope is essentially unchanged, 0.56 against 0.58, so including them would widen the calibration gap rather than narrow it. Each unit contributes a single observation at that date, which removes the within-unit serial correlation of the full panel and evaluates both forecasters at the same information point.

The human consensus is close to the Mincer–Zarnowitz calibration benchmark: its slopes

Figure 1: Reliability Diagrams Against Realized Outcomes

Alt text: Two square panels. Stated forecast probability runs left to right and the realized frequency of that outcome runs bottom to top, with a diagonal reference line for perfect calibration. The human curve lies close to the diagonal in both settings. The model curve is far flatter, sitting well below the diagonal where it states high probabilities and above it where it states low ones.



Notes: Each forecaster's stated probability is plotted against how often the predicted outcome actually occurs, so a forecaster on the 45° line is well calibrated. Panel A is prediction markets (the LLM persona crowd against the market price) and Panel B is analyst earnings (the informed EPS prompt against the analyst consensus). Because calibration requires one observation per event, the samples here are snapshots rather than the estimation panels used elsewhere. Panel A is the 8,131 markets of the canonical prediction-market panel at each market's last traded day. Panel B takes each firm-year's final analyst update, requires that update to fall within thirty days of the fiscal period end, and excludes fiscal year 2023 and observations missing the outcome, the consensus, or the LLM probability, leaving 2,351 firm-years; it does not impose the median-analyst, forecast-horizon, or sector requirements of the panel described in Section 2, so it is not a subset of the 2,649 firm-years reported there.

are $\hat{\beta} = 1.02$ and 1.00, and its probabilities explain 68% and 84% of the variation in realized outcomes. The joint restrictions $\alpha_f = 0$ and $\beta_f = 1$ are rejected in both settings ($F = 20.2$ and 6.4; $p < 0.001$ and $p = 0.0017$), but the departures are economically small relative to those for the LLM: a slope of 1.02 against 0.21, and 1.00 against 0.58. These restrictions test calibration, not the full set of conditions required for rational expectations. For the LLM, the joint null is rejected decisively ($F = 463$ and 952, $p \ll 0.001$): its slopes of 0.21 and 0.58 fall well below one, its intercepts of 0.15 and 0.36 exceed the human's, and its probabilities explain only 3.2% and 24% (Table 1). The reliability diagrams of Figure 1 show

Table 1: Forecast-Calibration Regressions (Mincer–Zarnowitz)

	N	$\hat{\alpha}$ $(H_0 : \alpha = 0)$	$\hat{\beta}$ $(H_0 : \beta = 1)$	R^2	MSE	F-test $(0, 1)$ (p-value)
Panel A: Prediction Markets						
LLaMA 3.3-70B	8,131	0.152*** (0.012)	0.207*** (0.026)	0.032	0.284	463.4 (< 0.001)
Human (market)	8,131	-0.013*** (0.002)	1.020*** (0.008)	0.683	0.057	20.2 (< 0.001)
Panel B: Analyst Earnings						
LLaMA 3.3-70B	2,221	0.357*** (0.019)	0.577*** (0.023)	0.236	0.210	952.0 (< 0.001)
Human (consensus)	2,221	0.016 (0.013)	0.996 (0.014)	0.842	0.035	6.4 (0.00172)

Notes: Mincer–Zarnowitz forecast-calibration regression $y_i = \alpha + \beta\,\hat{p}_i + \varepsilon_i$, where $y_i \in \{0, 1\}$ is the realized outcome and $\hat{p}_i$ the forecaster's stated probability; $\alpha = 0$ and $\beta = 1$ are joint calibration restrictions but are not sufficient for full rational expectations. The final column reports the F-statistic on the joint null $(\alpha, \beta) = (0, 1)$; stars on $\hat{\alpha}$ test $\alpha = 0$ and on $\hat{\beta}$ test $\beta = 1$. MSE is the mean squared forecast error (the Brier score for binary outcomes). Standard errors are clustered by evaluation date in Panel A and two-way by company and fiscal year in Panel B, with residual degrees of freedom as elsewhere in the paper. Snapshot dates as in Figure 1: the last pre-resolution day for prediction markets (8,131 markets) and each firm-year's final analyst update, required to fall within thirty days of the fiscal period end. The analyst sample additionally requires a GICS sector match, giving 2,221 firm-years against the 2,351 of Figure 1. Neither is a subset of the 2,649-firm-year panel of Section 2, which imposes median-analyst and forecast-horizon requirements that the calibration snapshot does not. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

the miscalibration: the human benchmark lies close to the 45° line in both settings, while the LLM's curve in prediction markets is nearly horizontal, varying little with the outcome; confident YES forecasts near 0.9 resolve YES only about 35% of the time. In analyst earnings the LLM's curve is closer to calibrated at the upper end of its range, consistent with its higher slope, but flattens at the bottom: a stated probability near 0.15 corresponds to a realized frequency near 0.60. A single affine recalibration cannot remove the nonlinear departures visible in the reliability curves or repair weak cross-event discrimination. The homo silicus premise does not require this benchmark to be unbiased, only that the LLM reproduce whatever biases traders or analysts carry.

The prediction-market reliability panel also provides a nonparametric calibration check on the human benchmark. Binning market prices by stated probability at each market's last traded day, the empirical YES frequency rises with price and spans the full $[0, 1]$ range, from 31.0% for contracts priced in $[0.40, 0.50)$, to 51.0% in $[0.50, 0.60)$, to 98.7% for those

priced at 0.90 or above, with close calibration at the extremes. Realized frequency exceeds mean price by 0.011 in the top decile and falls short of it by 0.006 in the bottom decile, and the ordering across deciles is monotone apart from a single reversal in the $[0.40, 0.50)$ decile, where realized frequency falls 14.6 percentage points short of mean price (31.0% against 0.456, $n = 300$). The wide cross-contract 25–75 percentile band in Figure 2 therefore reflects substantial heterogeneity in market-implied probabilities and realized frequencies, rather than only noise in the human benchmark.

3 The LLM–Human Gap

To compare across forecasts resolving to different outcomes, we align each forecaster's raw probability with the realized direction g_i, turning the probability of the favorable label into the probability of the *realized* outcome, so that a higher p_{it}^f reflects stronger directional alignment with it. For $f \in \{\text{LLM}, H\}$,

$$p_{it}^f = \begin{cases} \tilde{p}_{it}^f & \text{if } g_i = 1 \\ 1 - \tilde{p}_{it}^f & \text{if } g_i = 0 \end{cases}.$$

We summarize the comparison by the *signed LLM–human gap*

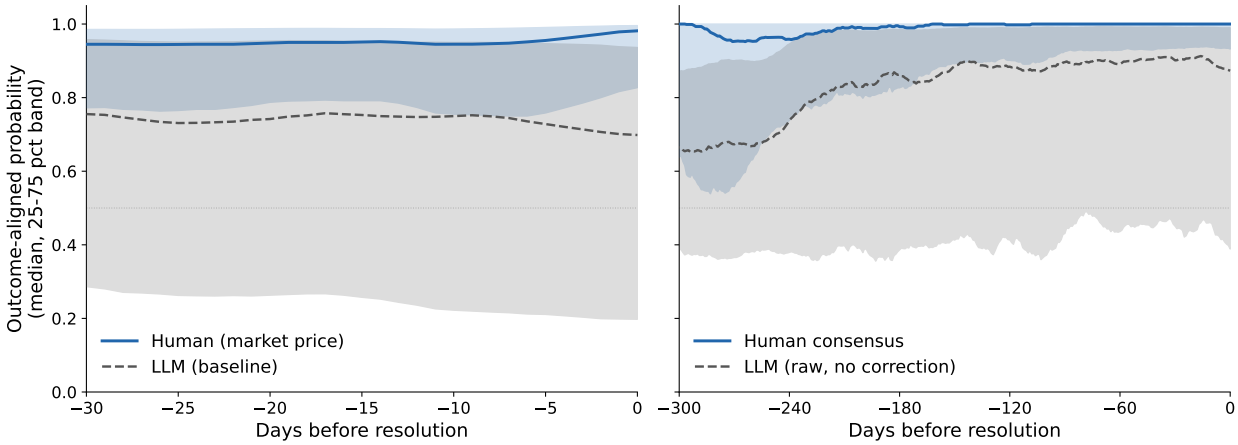
$$s_{it} = p_{it}^{\text{LLM}} - p_{it}^H,$$

where a negative s_{it} means the LLM assigns lower probability to the realized outcome than the human benchmark and a positive s_{it} means it assigns higher probability. Under the

homo silicus premise this gap should be zero on average and should not vary systematically with observable event characteristics.

Figure 2: Raw LLM–Human Gap in Prediction Markets and Analyst Earnings

Alt text: Two panels sharing a vertical axis of outcome-aligned probability, each drawn as a median line inside a shaded twenty-fifth to seventy-fifth percentile band, with a dotted line at one half. Time before resolution runs left to right, thirty days on the left and three hundred on the right. The human line sits near the top of both panels and stays roughly flat. The model line runs well below it, flat on the left and rising on the right without reaching it.



Notes: Outcome-aligned probability (median, 25–75 percentile bands) by trading day relative to resolution (prediction markets, left) and days before the firm-year's resolution date, its final analyst update (analyst earnings, right). Each panel plots the LLM at its per-setting homo-silicus baseline: in prediction markets (left) the demographic persona-crowd mean, with the median and 25–75 percentile band taken across markets at each trading day; in analyst earnings (right) the informed prompt supplying the firm's recent EPS. These are raw forecasts, before the training-corpus exposure correction of Sections 4.1–4.2. The plotted median and quartiles are centered rolling means over three trading days (prediction markets) and fourteen days (analyst earnings); smoothing the quartiles narrows the band slightly relative to the unsmoothed daily interquartile range.

Figure 2 shows that the gap is large in both settings: the mean gap is $\bar{s} \approx -23.4$ pp in prediction markets and -16.4 pp in analyst earnings over each setting's full window. The human benchmark stays close to the realized outcome throughout, averaging 0.840 in prediction markets and 0.838 in analyst earnings (medians 0.95 and 1.00). The LLM assigns far lower probability to that outcome across the entire horizon, averaging 0.606 in prediction markets and 0.674 in analyst earnings (medians 0.74 and 0.82). In prediction markets, the LLM's forecast is largely insensitive to the resolution approaching, so the gap

is widest near resolution, when the market is most informative. In analyst earnings, where the model conditions on accumulating quarterly EPS, its forecast converges partially toward the outcome, and the gap narrows modestly as resolution approaches.

3.1 Aggregation: The Persona Crowd

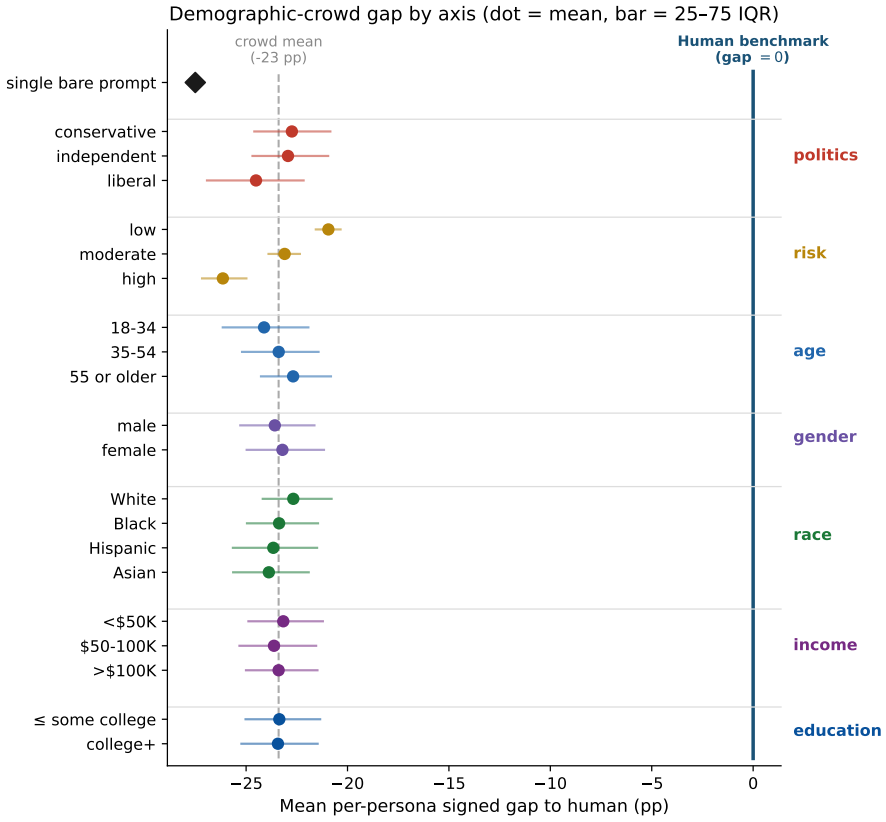
Aggregating the persona crowd does not reproduce the market: the crowd barely tracks the human price, with a correlation of only about 0.2, and its mean stays near 0.5 instead of following the market to its confident extremes. This compression does not arise from averaging over personas: the market price falls between the least and most confident personas in only 23% of markets (Appendix Figure IA1).

Figure 3 plots the demographic crowd's mean signed gap for each persona trait, with reference lines at the crowd mean and at zero, where the LLM matches the human benchmark. Almost every level sits on the crowd mean of about -23 pp. The single bare prompt plotted alongside the crowd uses the No/Yes label order of the persona elicitation and sits at -27.5 pp, about 4 pp closer to the benchmark than the Yes/No-ordered baseline of Section 5; the difference is the label-order effect documented there. The crowd's demographic mix need not match the platforms' actual trader population. Because the gaps vary little across the observed marginal trait levels, simple reweighting over those traits is unlikely to close the aggregate gap. The personas barely disagree: on a given market-date, their forecasts have a cross-persona standard deviation of only about 0.07. This spread stays nearly fixed as the outcome approaches, even as market prices become more concentrated near the realized outcome.

Some traits do move the forecast, consistent with this literature. Financial risk tolerance,

Figure 3: The Prediction-Market Gap Across the Demographic Persona Crowd

Alt text: A horizontal forest plot. Rows are grouped into seven blocks by trait, each block a different colour, with a marker for the mean and a bar for the interquartile range. A dashed vertical line marks the crowd mean near negative twenty-three percentage points and a solid line marks the human benchmark at zero. Almost every row sits on the dashed line, so the rows stack into a narrow column far from zero. A separate diamond at the top marks the single bare prompt, further left than any persona row.



Notes: Mean signed LLM–human gap (percentage points) by persona characteristic for LLaMA 3.3-70B, with reference lines at the human benchmark (gap = 0) and the crowd mean (−23 pp).

a driver of investor heterogeneity in Fedyk et al. (2024), shifts the gap about 5 pp from low to high, and political orientation, which Horton et al. (2023) and Bhagwat et al. (2026) find especially active, moves it about 2 pp from conservative to liberal. These trait effects are small relative to the 23 pp gap.

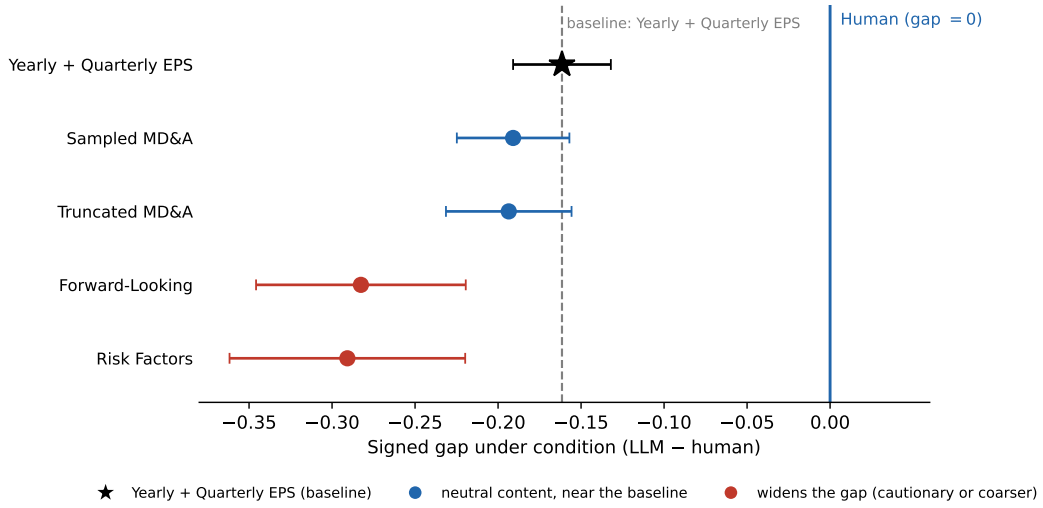
3.2 The Gap Across Context Conditions

A natural explanation for the LLM–human gap is an information gap: the model lacks what the human sees. We test this explanation by progressively enriching the prompt with the

data an analyst would use: starting from the analyst EPS baseline, we append narrative sections of the firm's mandatory SEC filings before the forecast question, in the spirit of retrieval-augmented generation (RAG).

Figure 4: The LLM–Human Gap Across Context Conditions

Alt text: A horizontal dot plot with one row per condition, each a marker with a confidence interval, measured in probability units. A dashed vertical line marks the baseline condition and a solid line marks the human benchmark at zero. A star marks the baseline row. The two neutral rows are blue and sit beside the dashed line; the two cautionary rows are red and sit further from zero, on the same side as the baseline.



Notes: Mean signed LLM–human gap under each context condition in analyst earnings, measured against the Yearly-plus-Quarterly EPS baseline (dashed) and the human benchmark (solid). Bars are 95% confidence intervals. Each condition's mean pools over outcome directions and exposure. The disclosure conditions do not use the firm's sector, so this figure applies the canonical filters of Section 2 without the Compustat sector requirement, giving 2,789 firm-years: the 2,649 reported there plus the 140 carrying no GICS code.

At the Yearly-plus-Quarterly EPS baseline, a gap of approximately −16 pp remains between the model and the analyst. The disclosure filings widen it further: the sampled and truncated MD&A excerpts by about 3 pp each, the more cautionary Risk Factors and Forward-Looking Statements by about 13 and 12 pp, and none of these conditions reaches the human benchmark.

None of the disclosure conditions closes the gap, and the cautionary sections move the forecast further from the benchmark. The supplied excerpts do not span the full information

set of an analyst, which can also include earnings calls, competitor filings, and accumulated sector expertise. More generally, researchers rarely observe the complete information set underlying a human belief. Adding context may reduce an information gap, but it cannot by itself validate the model as a belief proxy.

4 The Structure of the Gap

The gap is the difference between the LLM's forecast and the human belief it is meant to measure. This matters once LLM-generated measures are used in downstream economic tasks, where such a difference can bias the results. Rather than treat the gap as a single number, we distinguish two systematic components tied to the model, its training corpus, and the elicitation task: training-corpus exposure and a directional prior.

4.1 Training-Corpus Exposure

The LLM's training data ends at a cutoff that is typically reported as a range of dates, a window $[T_0, T_1]$ (December 2023 in LLaMA 3.3-70B), so outcomes resolving before T_0 may already appear in the training corpus while those resolving after T_1 do not. We partition forecasts by exposure status,

$$d_i = \begin{cases} 1 & \text{if } \tau_i < T_0 \text{ (exposed)}, \\ 0 & \text{if } \tau_i > T_1 \text{ (unexposed)}, \end{cases}$$

where τ_i is the resolution date of forecast i; we exclude forecasts resolving inside the window, where exposure is ambiguous, from the estimation sample. This ambiguous window is the

cutoff month in prediction markets and all of fiscal year 2023 in analyst earnings: for a fiscal-2023 firm-year, many analyst-update dates precede the cutoff, so interim discussion of the year can appear in the training corpus even though the annual outcome resolves after the window closes, making exposure partial.

Evaluating the LLM on those exposed events therefore conflates memorization with genuine forecasting. Splitting the sample at the cutoff shows that the gap is narrower on the exposed side: -18.9 pp against -23.9 pp unexposed in prediction markets, and -14.2 pp against -20.0 pp in analyst earnings (exposed events are 815 of the 8,131 prediction-market contracts and 1,656 of the 2,649 analyst firm-years). Pooling the narrower exposed events with the unexposed ones therefore understates the deployment-relevant gap: in actual use every event resolves after training, so the post-cutoff gap is the one a user would face.

The exposure classification is binary, but its effect on a continuous elicited probability is measured in percentage points rather than only through changes in the most likely answer. We estimate the magnitude of training-corpus exposure, $\hat{\lambda}$, by comparing signed gaps across the December 2023 boundary, holding the forecasting rule fixed and controlling for task difficulty:

$$s_{it} = \alpha + \lambda d_i + \Gamma_{it}'\eta + \varepsilon_{it}.$$

The controls Γ_{it} collect observable task characteristics: the market's platform, days to resolution, and trading volume in prediction markets; analyst coverage, firm fundamentals, and sector fixed effects in analyst earnings. For d_i to identify the causal effect of exposure, potential signed gaps must be conditionally independent of cutoff-side assignment given Γ_{it}, with overlap in the support of those covariates. The reported cutoff also must create a one-sided

contrast: forecasts resolving after the cutoff window cannot contain outcome-resolving text in the training corpus under the stated convention. Under these assumptions, the conditional cross-cutoff difference identifies the exposure effect. The design does not require uniform training-data density before T_0, but imperfect correspondence between pre-cutoff eligibility and actual corpus inclusion makes λ an intention-to-expose contrast. We cluster standard errors two-way by market and evaluation date in prediction markets and by company and fiscal year in analyst earnings, and this convention applies throughout the paper unless otherwise noted. The analyst panel contains only five fiscal-year clusters, so the analyst standard errors are effectively determined by that dimension. We therefore interpret threshold-based inference along it cautiously and emphasize coefficient magnitudes and stability.

Table 2 reports these estimates with controls added progressively from column (1) to column (4). A positive estimate of $\hat{\lambda}$ means the exposed forecasts sit closer to the realized outcome, consistent with training-corpus exposure narrowing the gap by adding post-resolution information. The narrowing survives the full controls in both settings: 4.0 pp in prediction markets and 5.4 pp in analyst earnings.[8] The exposed analyst years (fiscal 2020–2022) overlap the COVID period, raising the concern that the narrowing reflects the pandemic rather than training exposure, but a post-COVID contrast (fiscal 2022 against 2024–2025) still narrows the gap, by +3.4 pp against +5.8 pp on the full split. The distance specifications are consistent with a discrete shift at the cutoff rather than a gradient on either side. Gradual exposure would make the narrowing grow with an outcome's distance before the cutoff, and a growing human information advantage over the fixed-cutoff model would make the gap

[8] Appendix C reports donut, placebo-cutoff, human-difficulty-placebo, and macro-regime specifications that support the training-corpus-exposure interpretation.

widen with distance after it; instead the estimated slope is flat before the cutoff and slightly positive after it (Appendix B.5).

Table 2: Baseline Training-Corpus Exposure in Prediction Markets and Analyst Earnings (Signed Gap)

	(1)	(2)	(3)	(4)
Panel A: Prediction Markets				
Exposure	0.0503***	0.0615***	0.0417**	0.0403**
	(0.016)	(0.017)	(0.017)	(0.017)
Source FE (Polymarket/Kalshi)	No	Yes	Yes	Yes
Market controls (log volume)	No	No	Yes	Yes
Calendar FE (weeks-to-resolution)	No	No	No	Yes
Observations	158,814	158,814	158,410	158,410
Clusters (markets)	8,131	8,131	8,131	8,131
R^2	0.001	0.002	0.005	0.008
Panel B: Analyst Earnings				
Exposure	0.0578***	0.0573***	0.0552***	0.0544***
	(0.016)	(0.017)	(0.020)	(0.021)
Analyst Coverage	No	Yes	Yes	Yes
Firm controls (log assets, ROA, book-to-market, cash ratio)	No	No	Yes	Yes
GICS sector FE	No	No	Yes	Yes
Calendar FE (weeks-to-resolution)	No	No	No	Yes
Observations	132,957	132,957	132,580	132,580
R^2	0.006	0.006	0.022	0.026

Notes: Exposure coefficient on the signed LLM–human gap (Panel A: prediction markets; Panel B: analyst earnings), with controls added progressively across columns. A positive coefficient means the gap is narrower for pre-cutoff observations. Standard errors are two-way clustered by market and evaluation date in Panel A and two-way clustered by company and fiscal year in Panel B. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

The $\hat{\lambda}$ estimated above is a single average across exposed forecasts, but exposure is not uniform across events: a sample concentrated in high-exposure events carries more bias than this average implies. We therefore let the exposure effect vary along an observable dimension Z_{it} by adding Z_{it} and its interaction with exposure:

$$s_{it} = \alpha + \lambda d_i + Z_{it}'\chi + (d_i \times Z_{it})'\pi + \Gamma_{it}'\eta + \varepsilon_{it}, \tag{2}$$

where χ absorbs baseline differences in the gap along Z_{it} and $(d_i \times Z_{it})'\pi$ identifies how the exposure effect varies with Z_{it}.

We first instantiate Z_{it} in Equation 2 as indicators for each setting's heterogeneity: the prediction-market topic category or the firm's GICS sector, writing π_z for the element of π on the indicator of cell z.[9] Figure 5 plots each category's net exposure effect, $\hat{\lambda} + \hat{\pi}_z$ (or $\hat{\lambda}$ for the omitted category, Politics in prediction markets and Industrials in analyst earnings), positive where exposure narrows that category's gap.[10] In prediction markets the narrowing concentrates in a few categories: Economics and Science & Technology are largest, both around 22 pp; Sports and Politics are moderate (9 and 5 pp, neither significant at the 5% level); and Crypto, the combined financial category (Financials, Commodities & Companies), Climate & Weather, and Entertainment show no detectable effect. In analyst earnings the narrowing concentrates in Energy (29 pp), Consumer Discretionary (11 pp), and Real Estate (10 pp), while Consumer Staples is positive but imprecisely estimated, and the remaining sectors are statistically indistinguishable from zero.

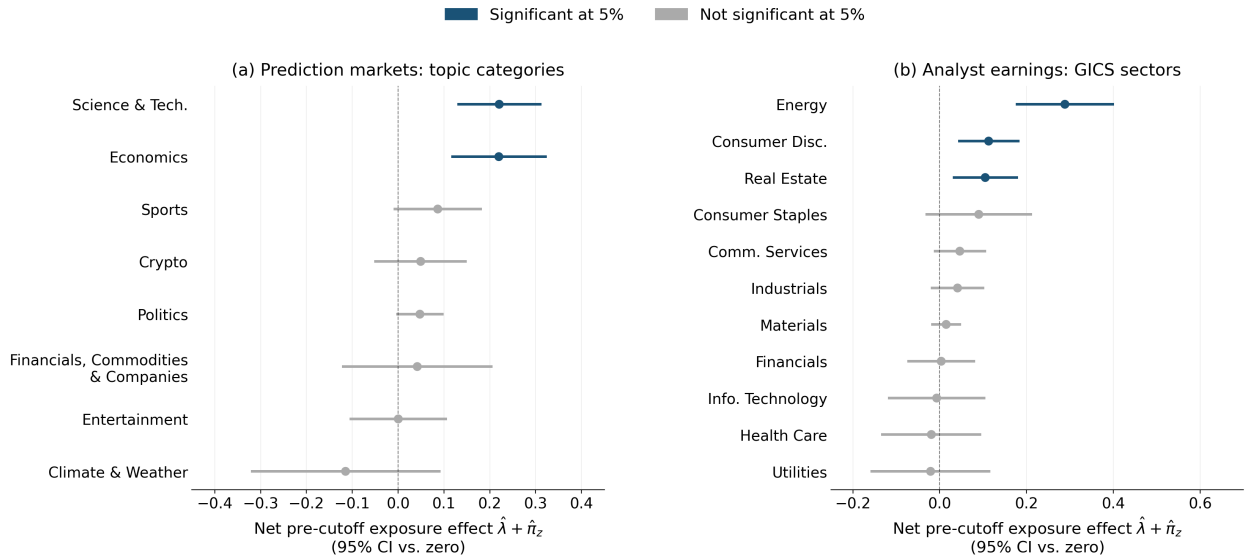
With the exposure evidence assembled, we can assess the exogeneity assumption of Section 4.1, which can fail in three main ways. First, events resolving before and after the cutoff may differ in composition or inherent difficulty. The two groups differ on some observables, which the regressions control for, while the human benchmark supplies a placebo that runs opposite to the principal difficulty threat: the outcome-aligned human probability is lower before the cutoff (Appendix B.2). Second, the calendar is not randomly assigned: the pre-cutoff analyst window coincides with the COVID years, so a macro-regime difference in earnings predictability could masquerade as exposure. Appendix C.1 decomposes the effect by fiscal year and re-estimates it without the COVID years. Third, the gap could drift

[9] Full regression estimates are in Appendix Tables IA10 and IA11.

[10] The net effect $\hat{\lambda} + \hat{\pi}_z$ does not depend on which cell is omitted. The figure and Appendix Table IA11 omit Industrials, while Table 4 omits Energy, so the $\hat{\lambda}$ reported in that table is the exposure effect for a different reference sector than the one plotted here.

Figure 5: Net Pre-Cutoff Exposure Effect by Topic Category and by Sector

Alt text: Two panels of horizontal intervals, prediction-market topics on the left and industry sectors on the right, each category on its own row and sorted by size. A vertical line marks zero. Rows whose interval clears zero are drawn in blue and the rest in grey; two rows are blue on the left and three on the right.



Notes: Each point is a category's net training-corpus exposure effect on the signed gap, $\hat{\lambda} + \hat{\pi}_z$, with 95% confidence intervals, for prediction-market topic categories in Panel (a) and GICS sectors in Panel (b). Blue marks categories with detectable exposure and gray marks those without, including prediction-market categories with too few pre-cutoff markets to estimate.

smoothly in calendar time for reasons unrelated to training, and the cutoff contrast would absorb that drift. The estimates are robust to donut windows around the boundary and to placebo cutoffs placed away from it, and the distance specification is consistent with a level shift at the cutoff rather than a smooth gradient (Appendix B).

4.2 The Directional Prior and Outcome Asymmetry

A training corpus that disproportionately documents one type of outcome can induce an event-independent tilt toward that outcome. Table 3 shows this direction-dependence. In the raw split the gap concentrates on NO-resolving contracts (−26.2 against −13.2 pp) and on EPS-increasing firms (−21.7 against −6.5 pp); removing the per-direction training-corpus exposure leaves it narrower on the favorable direction in both settings (−15.6 against −26.0

pp on YES- versus NO-resolving contracts, −17.9 against −24.1 pp on EPS-increasing versus EPS-decreasing firms), even flipping which side is larger in analyst earnings.

Table 3: LLM–Human Gap by Realized Outcome

	Prediction Markets			**Analyst Earnings**		
	LLM	Market	Gap	LLM	Consensus	Gap
Overall (raw)	60.6	84.0	−23.4	67.4	83.8	−16.4
Overall (exposure-adjusted)	60.3	84.0	−23.8	63.7	83.8	−20.1
Favorable (Yes / Increase), raw	52.9	66.0	−13.2	66.7	88.4	−21.7
Favorable, exposure-adjusted	50.4	66.0	−15.6	70.4	88.4	−17.9
Unfavorable (No / Decrease), raw	62.8	88.9	−26.2	68.7	75.2	−6.5
Unfavorable, exposure-adjusted	62.9	88.9	−26.0	51.0	75.2	−24.1

Notes: LLM and human outcome-aligned probabilities (in percent) and their signed difference ("Gap," in percentage points) by realized outcome direction. 23.5% of the 8,131 prediction-market contracts resolve YES; 67.8% of the 2,649 analyst firm-years are EPS-increasing. Exposure-adjusted rows subtract the column (1) exposure terms of Table 4 (the pooled-sample averages) from each exposed observation: $p_{it}^{\text{LLM,adj}} = p_{it}^{\text{LLM}} - \hat{\lambda} d_i - \hat{\theta}\,(d_i \times g_i)$.

The directional regression separates the prior from the exposure asymmetry by adding g_i and its exposure interaction to Equation 2:

$$s_{it} = \alpha + \lambda d_i + \gamma g_i + \theta\,(d_i \times g_i) + Z_{it}'\chi + (d_i \times Z_{it})'\pi + \Gamma_{it}'\eta + \varepsilon_{it}. \tag{3}$$

We do not carry the raw-and-adjusted distinction of Table 3 forward: every later specification retains the exposure term d_i alongside the realized-outcome indicator g_i and their interaction, so the gap estimates that follow are net of training-corpus exposure.

The gap depends on outcome direction in both settings, but through different channels (Table 4). In prediction markets (Panel A), the channel is a directional prior γ, the model's tilt toward the favorable label beyond the human benchmark's own direction-dependence: $\hat{\gamma} = 0.117$ in the fully-controlled column (4), so the gap is about 11.7 pp higher on YES- than NO-resolving markets even where the model has not seen the outcome, and stable as controls are added (0.104 to 0.117). Exposure adds a direction-asymmetry θ, the amount it narrows

the gap on one direction relative to the other: it does not narrow NO-resolving markets ($\hat{\lambda} = -0.023$, column (1)) but narrows the YES-resolving ones by 16.7 pp ($\hat{\lambda} + \hat{\theta} = +0.167$), with $\hat{\theta}$ between +0.177 and +0.190 across all four columns.

Table 4: Directional Bias in Prediction Markets and Analyst Earnings (Signed Gap)

	(1)	(2)	(3)	(4)
Panel A: Prediction Markets				
Exposure	−0.0232	−0.0248	0.0007	−0.0010
	(0.019)	(0.020)	(0.026)	(0.026)
Outcome	0.1036***	0.1120***	0.1165***	0.1167***
	(0.014)	(0.014)	(0.014)	(0.014)
Exposure × Outcome	0.1901***	0.1895***	0.1774***	0.1787***
	(0.033)	(0.033)	(0.033)	(0.033)
Source FE, Log Volume	No	Yes	Yes	Yes
Exposure × Topic interactions	No	No	Yes	Yes
Calendar FE (weeks-to-resolution)	No	No	No	Yes
Observations	158,814	158,410	158,410	158,410
Clusters (markets)	8,131	8,131	8,131	8,131
R^2	0.024	0.030	0.079	0.083
Panel B: Analyst Earnings				
Exposure	0.2720***	0.2738***	0.4814***	0.4800***
	(0.077)	(0.079)	(0.101)	(0.099)
Outcome	0.0620*	0.0622*	0.0379	0.0377
	(0.037)	(0.036)	(0.041)	(0.041)
Exposure × Outcome	−0.3321**	−0.3359**	−0.3110**	−0.3106**
	(0.163)	(0.165)	(0.153)	(0.153)
Firm controls (log assets, ROA, book-to-market, cash)	No	Yes	Yes	Yes
Exposure × GICS Sector interactions	No	No	Yes	Yes
Calendar FE (weeks-to-resolution)	No	No	No	Yes
Observations	132,957	132,580	132,580	132,580
R^2	0.083	0.085	0.106	0.110

Notes: The dependent variable is the signed LLM–human gap. Without heterogeneity interactions, λ is the exposure effect for unfavorable-outcome observations; γ is the directional prior in the unexposed regime, and θ is the exposure asymmetry. In Panel A, columns (2)–(4) add source and log-volume controls. In Panel B, columns (2)–(4) add firm controls. Column (3) also adds exposure × topic (A) or exposure × sector (B) interactions, so λ is then the effect for the reference cell (NO-resolving Politics markets in A; EPS-decreasing Energy firms in B), which is why Panel B's column-(3) exposure coefficient rises to 0.47. Column (4) adds event-time fixed effects. SEs two-way clustered by market and date (A), company and fiscal year (B). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

In analyst earnings (Panel B), the channel is exposure, not a prior. The prior is small and insignificant (+3.8 pp under full controls; both +6.2 pp in columns (1)–(2), significant at the 10% level in both). The exposure asymmetry is large and points the opposite way from prediction markets: exposure narrows the gap by 27.2 pp on EPS-decreasing firms

($\hat{\lambda} = +0.272$, column (1)) but not on EPS-increasing ones ($\hat{\lambda} + \hat{\theta} = -0.060$), an asymmetry $\hat{\theta} = -0.332$ stable across columns (-0.311 to -0.336). As with topics, the asymmetry falls on the outcomes that generate distinctive post-resolution coverage: EPS-decreasing firm-years in analyst earnings, and YES-resolving contracts in prediction markets. Figure 6 plots these corrections by direction; each direction's contribution to the pooled gap scales with its sample share, so a large narrowing in a small cell moves the overall gap only modestly.

A possible explanation for the directional prior is that the LLM has learned the frequency of outcomes appearing in training. Under frequency-matching the model would state the base rate q of the favorable outcome.[11] With only 23.5% of prediction-market contracts resolving YES and 67.8% EPS-increasing firm-years in analyst earnings (Table 3), frequency-matching predicts a prior of -30.1 pp in prediction markets and $+22.3$ pp in analyst earnings. We instead observe $+11.7$ pp on the crowd baseline in prediction markets, the opposite sign, and only about $+4$ pp in analyst earnings, well below the prediction. Simple frequency matching therefore does not explain the directional prior. The pattern points to a learned directional mapping with several candidate sources: the composition of the corpus, the favorable connotations answer labels accumulate in training, and post-training toward answers raters prefer.

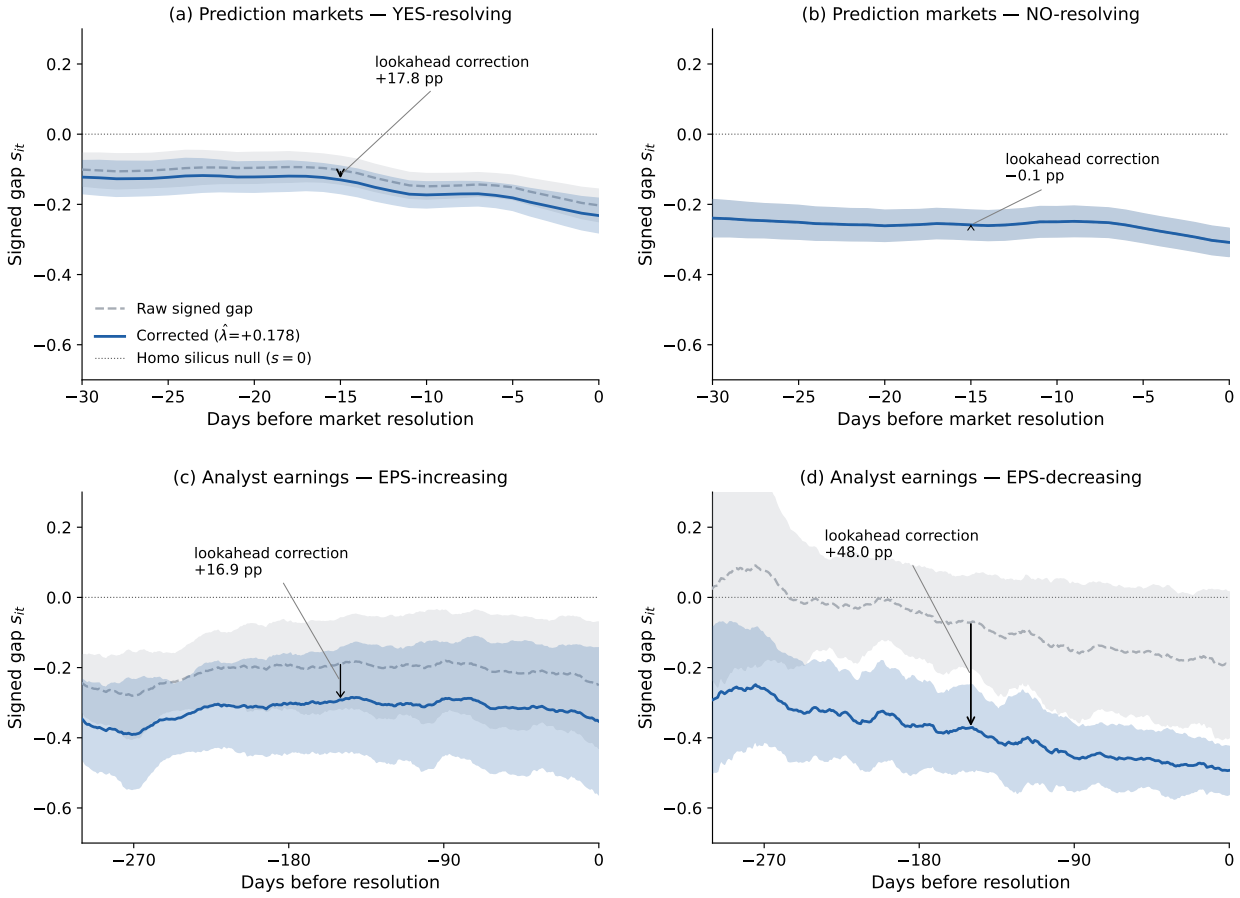
A caution follows for research that uses these probabilities: the directional prior cannot be removed from a live forecast, since netting it out requires the realized outcome, which is unknown at forecast time. It can only be estimated and reported as a diagnostic after

[11] The implied prior is then $\gamma = 2q - 1 - (p_F^H - p_U^H)$, where the second term corrects for the human benchmark's own level difference across directions: if $p_{it}^{\text{LLM}} = q$ for every event, the model's outcome-aligned probability is q on favorable events and $1 - q$ on unfavorable ones, so the cross-direction difference in the signed gap is $[q - p_F^H] - [(1 - q) - p_U^H] = 2q - 1 - (p_F^H - p_U^H)$, with p_F^H and p_U^H the human outcome-aligned means by direction from Table 3. The conclusion is unchanged under the simpler benchmark $2q - 1$ that omits the human correction (-53.0 and $+35.5$ pp).

outcomes resolve.

Figure 6: Directional Asymmetry of Training-Corpus Exposure

Alt text: Four panels in two rows, splitting each setting by which side the outcome resolved on. Time before resolution runs left to right and the signed gap runs bottom to top, with a dotted line at zero. A grey dashed line shows the raw gap and a blue solid line the corrected gap, each inside a shaded band. A labelled arrow in every panel marks the size of the correction, which is large in three panels and near zero in the second.



Notes: Mean signed gap by event time relative to resolution (trading days in prediction markets, weeks in analyst earnings), shown raw (gray, dashed) and after correcting for training exposure separately within each outcome direction (blue, solid), with the labeled arrows giving the per-direction correction. The top row is prediction markets (YES-resolving and NO-resolving) and the bottom row is analyst earnings (EPS-increasing and EPS-decreasing). Shaded bands are ±1.96 standard errors. Series are smoothed as in Figure 2, by centered rolling means over three trading days (prediction markets) and fourteen days (analyst earnings).

4.3 What the Gap's Components Can and Cannot Remove

We now estimate the decomposition out of sample, from two vantage points. A researcher correcting a forecast before the firm-year resolves can use only what is known at elicitation,

the firm's exposure, sector, and forecast horizon among them. A researcher correcting it after resolution can also use the realized outcome g_i. We call these the ex ante and ex post cases. The ex ante gap model carries no g_i terms. The ex post model adds $\gamma g_i + \theta\,(d_i \times g_i)$ to it and is re-estimated, so it bounds what the decomposition can explain in hindsight rather than what a forecaster could have removed. Both are estimated on the raw increase-probability scale. The ex ante case has no alternative, since the outcome-aligned transform is a function of g_i; the ex post case uses the same scale so that the outcome terms are the only difference between them. Of each correction we then ask two questions: whether it removed the structure it was built from, on the scale it was built on, and what it does to the outcome-aligned gap that Sections 4.1 and 4.2 decompose.

We split the firms into two folds, assigning each firm so that all of its observations fall in the same fold. On one fold we estimate

$$\tilde{s}_{it} = \alpha + \lambda d_i + Z_i'\chi + (d_i \times Z_i)'\pi + \Gamma_{it}'\eta + e_{it}, \tag{4}$$

where $\tilde{s}_{it} = \tilde{p}_{it}^{\mathrm{LLM}} - \tilde{p}_{it}^{H}$. Equation (4) is the ex ante model, and it drops the outcome terms of Equation (3) rather than fitting them and setting them aside. We subtract the fitted gap from the raw LLM probability on the held-out fold and realign,

$$\tilde{p}_{it}^{C} = \tilde{p}_{it}^{\mathrm{LLM}} - \hat{\tilde{s}}_{it}^{(-k)}, \qquad s_{it}^{C} = p_{it}^{C} - p_{it}^{H}, \tag{5}$$

where $\hat{\tilde{s}}_{it}^{(-k)}$ is the fitted gap from the fold that excludes firm i. Applying Equation (5) gives an *ex ante linear corrected score*; applying the ex post model gives its ex post counterpart.

We retain the notation $\tilde{p}^C$ to display the subtraction, but do not call the result a probability unless it is mapped back to $[0, 1]$.[12] Because the fitted gap comes from Equation (4), which carries sector terms, the amount subtracted varies by sector and is estimated on other firms. This is the out-of-sample counterpart of the exposure adjustment $p_{it}^{\text{LLM}} - \hat{\lambda} d_i - \hat{\theta}\,(d_i \times g_i)$ used earlier in this section, which subtracts pooled-sample averages. We then estimate on the second fold and correct the first, so the two sets of held-out corrections together cover the full panel. We then re-estimate Equation (3) on the corrected aligned gap s_{it}^C, which uses the realized outcome, as any evaluation against that outcome must, even when the correction being evaluated does not.

Table 5 reports the result. The uncorrected row reproduces the fully controlled estimates reported earlier in this section, with $\hat{\lambda} = 0.480$, $\hat{\gamma} = 0.038$, and $\hat{\theta} = -0.311$ on a gap of -16.3 percentage points, so the held-out exercise decomposes the same structure documented on the full sample. In the ex ante case the gap narrows to -12.6 percentage points using exposure alone and to -12.0 once sector heterogeneity is added. That the pooled gap narrows here, while the exposure adjustment used earlier in this section widens it, is not evidence that lookahead has been removed. An ex ante correction can remove an average exposure effect, and with sector terms a heterogeneous one, but lookahead runs through the outcome direction. Before any correction, exposure nearly closes the gap on unfavorable outcomes, from -24.1 percentage points when unexposed to $+3.0$ when exposed, and widens it on favorable outcomes, from -17.8 to -23.8. A correction that cannot use g_i fits one coefficient for both directions and leaves that asymmetry in place, and the narrowing comes instead

[12] The linear subtraction *falls outside* $[0, 1]$ for 28.2% of observations (26.8% above one and 1.4% below zero), confirming that it is a regression score rather than a probability. Appendix Table IA9 reports clipped and log-odds alternatives that remain bounded.

from the average level difference it removes, which moves from −11.7 to +0.5 percentage points on the raw scale. Exposure falls from 0.480 to 0.226, while the directional prior rises from 0.038 to 0.117. The gap still differs across industries by nearly as much as before: the standard deviation of the sector effects falls only from 6.6 to 5.4 percentage points. The exposure asymmetry also falls, from −0.311 to −0.050, which follows from the mapping between the two scales rather than from the ex ante correction reaching an outcome-defined term.

In the ex post case the gap falls to −0.1 percentage points using exposure alone and to −0.2 once sector heterogeneity is added. The directional coefficients are economically small, and the reference-sector exposure coefficient falls to 0.129. Because these corrections use the realized outcome, they bound what the decomposition accounts for rather than supplying an adjustment available before resolution.

The held-out loss functions give the key qualification. On the raw increase-probability scale, the ex ante linear correction removes the mean bias (−0.117 to +0.005) but does not lower mean absolute error (0.283 to 0.283), and its held-out R^2 is only 0.051. Clipping reduces mean absolute error to 0.253, while a bounded log-odds correction reduces it to 0.223; the ex post clipped benchmark reaches 0.212 only by using g_i. Appendix Table IA9 reports the complete comparison.

The distinction between what the correction removes out of sample and what remains gives a practical map of the gap. Exposure, sector, and horizon are observable when the probability is elicited and can be subtracted in advance. They remove average bias, but explain little held-out individual variation. The directional prior cannot be subtracted in advance, because it is defined relative to the outcome. A researcher working with an elicited

Table 5: Held-Out Decomposition of the Linear Corrected Score

Correction model	Mean $\bar{s}$ (pp)	α	λ	γ	θ	Sector s.d.	Held-out R^2
Uncorrected	−16.3	-0.493^{***}	$+0.480^{***}$	+0.038	-0.311^{**}	0.066	
		(0.069)	(0.099)	(0.041)	(0.153)		
Ex ante: no outcome terms, formed ex ante							
Exposure only	−12.6	-0.506^{***}	$+0.335^{***}$	$+0.125^{***}$	−0.064	0.066	0.033
		(0.072)	(0.095)	(0.040)	(0.150)		
With sector heterogeneity	−12.0	-0.398^{***}	$+0.226^{*}$	$+0.117^{**}$	−0.050	0.054	0.051
		(0.068)	(0.122)	(0.046)	(0.153)		
Ex post: adds γg_i *and* $\theta\,(d_i \times g_i)$							
Exposure only	−0.1	-0.233^{***}	$+0.185^{**}$	−0.013	+0.018	0.064	0.168
		(0.085)	(0.086)	(0.042)	(0.150)		
With sector heterogeneity	−0.2	-0.201^{**}	+0.129	−0.003	+0.015	0.050	0.183
		(0.081)	(0.121)	(0.039)	(0.152)		

Notes: The dependent variable is the outcome-aligned LLM–human gap left by each correction; Equation (3) is estimated on it over the full sample, and only the correction is estimated out of sample. The coefficients therefore measure the structure *remaining* after each correction, not the coefficients used to build it, which are estimated within the training fold on the raw-scale gap and are not shown. "Exposure only" fits the raw-scale gap on an intercept, exposure d_i, firm controls, and forecast-week effects; "with sector heterogeneity" adds sector indicators and exposure × sector interactions, giving Equation (4). The ex post rows add $\gamma g_i + \theta\,(d_i \times g_i)$ and are re-estimated. $\bar{s}$ is the mean gap in percentage points, sector s.d. the standard deviation of the eleven sector effects, and held-out R^2 the correction's fit for the raw-scale gap on excluded firms. Standard errors in parentheses are two-way clustered by firm and fiscal year and treat the held-out correction as given. They therefore do not incorporate first-stage estimation uncertainty; with only five fiscal-year clusters, the stars are descriptive. The held-out loss measures in Appendix Table IA9 are the primary validation of the correction. Sample: 132,580 firm-year–date observations, 578 companies, 2,644 firm-years, the 2,649 of Section 2 less the five firm-years of one dual-listed company for which Compustat reports no market value. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

probability can therefore adjust for the observable structure in advance, but the prior has to be validated against human data rather than corrected. The gap is therefore not an artifact of what a researcher can observe: most of it survives that adjustment, and what survives is defined relative to the outcome.

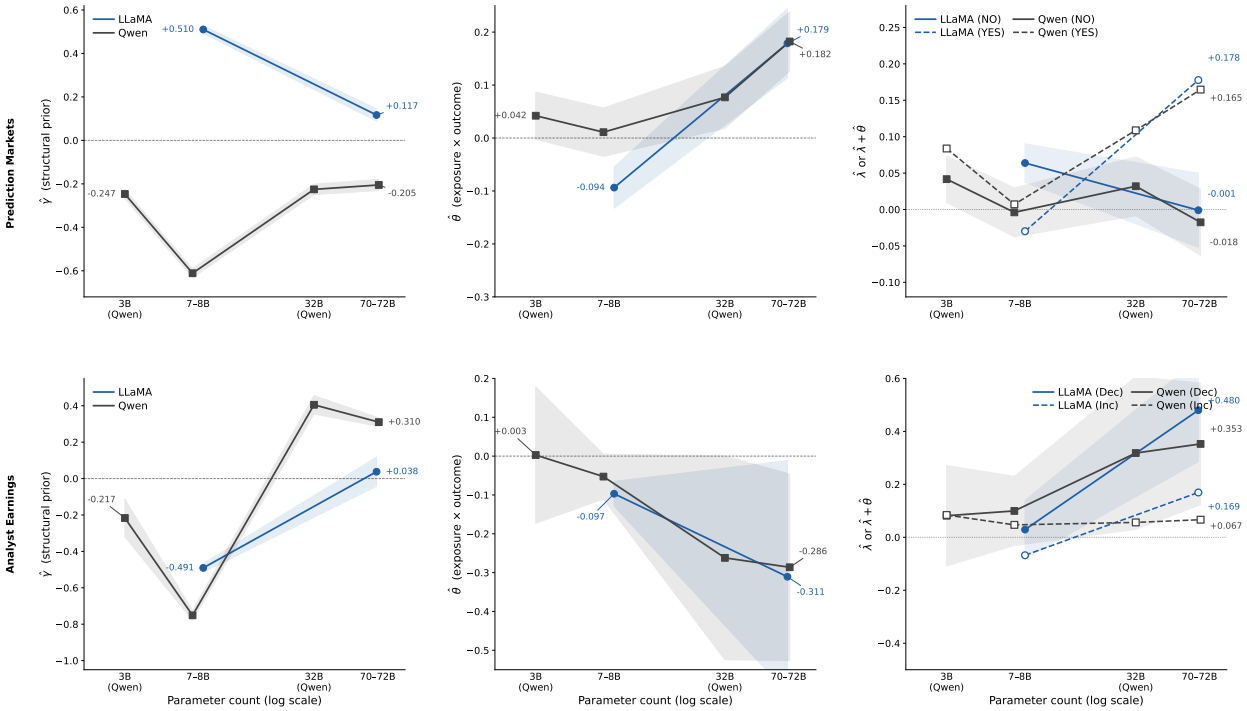
4.4 Scale and Model Family

The two components so far describe a single model, and whether they carry over to other models is itself part of the structure. There is no presumption that scale removes them: a larger model may use context more effectively, but it may also encode more training-era information or weight the same response labels differently. Size comparisons within and

across families are confounded by differences in data, post-training, and tokenization, so we estimate the components by model. For each model, we estimate the structural directional prior $\hat{\gamma}$, the directional asymmetry of exposure $\hat{\theta}$, and pre-cutoff exposure $\hat{\lambda}$ under Equation 3, using that model's exposure-classification convention and the paper's informed baselines, the persona crowd in prediction markets and the EPS prompt in analyst earnings.[13]

Figure 7: Cross-Model Coefficient Estimates: LLaMA 3.x and Qwen 2.5

Alt text: Six panels in two rows, prediction markets above and analyst earnings below. Model size runs left to right on a logarithmic scale from three billion to seventy-two billion parameters. Each panel plots one estimate, with a blue line for the LLaMA family and a grey line for the Qwen family inside shaded bands. In the leftmost panels the two families sit on opposite sides of the zero line, so the directional prior does not share a sign across families or converge as models grow.



Notes: The estimated directional prior, exposure asymmetry, and training-corpus exposure by model size, for LLaMA 3.x (blue) and Qwen 2.5 (gray), with prediction markets in the top row and analyst earnings in the bottom row. Both settings use the paper's baseline elicitation, and the exposure effect is shown separately by outcome direction. Bands are 95% confidence intervals.

Figure 7 plots the three coefficients across model size in both settings. Larger models do not move the components toward zero, and no monotone scaling pattern appears. The

[13]The grid spans LLaMA 3.1-8B, LLaMA 3.3-70B, and Qwen 2.5 at 3B, 7B, 32B, and 72B (Figure 7); the LLaMA models use their reported cutoff, while the Qwen models use the conservative classification window described in Appendix Table IA1.

structural prior is positive for both LLaMA models in prediction markets, but in analyst earnings it is near zero for the 70B and strongly negative for the 8B; for Qwen it is negative in prediction markets and flips sign across sizes in analyst earnings. At the largest models, the exposure asymmetry is positive in prediction markets and negative in analyst earnings (Appendix D.3). A researcher cannot assume an arbitrary LLM is free of these components or that estimates from one model carry over to another, because the components recur across families and sizes without the same signs, magnitudes, or significance. These estimates therefore constitute a model- and setting-specific validation exercise: recovering the components requires model outputs, realized outcomes, a cutoff-based exposure classification, and a human benchmark.

5 Forecast Responses to Prompt Form and Supplied Signals

The components so far are systematic features of the model–task interaction. We now hold the forecasting question fixed (the same event, resolving on the same date) and vary the prompt around it, both what it supplies and how it is posed: the text of a disclosure, the human's own probability, deliberately incorrect numbers, the order and wording of the response labels, and instructions to reason. Together these measure how far the continuation probability tracks a belief rather than the surface features of the prompt.

5.1 Supplying Firm Disclosures

In the analyst setting, we analyze four disclosure conditions: a sampled MD&A excerpt, a truncated MD&A excerpt, the MD&A's Forward-Looking Statements, and the Risk Factors.

The disclosures are text, so they enter the forecast through two channels: the earnings signal they contain and the learned associations of their tokens.[14] We stack them against the Yearly-plus-Quarterly EPS baseline over the S&P 500 analyst firm-years and estimate the signed gap:

$$s_{it,k} = \alpha + \sum_k \beta_k \mathcal{A}^{(k)} + \lambda\, d_i + \sum_k \mu_k \left(d_i \times \mathcal{A}^{(k)}\right) + \sum_k \nu_k \left(g_i \times \mathcal{A}^{(k)}\right)$$
$$+ \gamma\, g_i + \theta \left(d_i \times g_i\right) + Z_{it}'\chi + \left(d_i \times Z_{it}\right)'\pi + \Gamma_{it}'\eta + \varepsilon_{it,k} \tag{6}$$

where the stacked sample has one row per firm-year-date it and condition k, $s_{it,k}$ is the signed gap under condition k, and $\mathcal{A}^{(k)}$ is an indicator for condition k, with the baseline as the omitted category, so $\hat{\beta}_k$ measures its effect on the gap. Each condition's effect on the gap is $\hat{\beta}_k$ on EPS-decreasing firms and $\hat{\beta}_k + \hat{\nu}_k$ on EPS-increasing ones, so the interaction $\hat{\nu}_k$ measures how that effect differs by outcome direction. Table 6 reports these estimates.

Starting from the informed baseline, we see a gap of about 16 pp. If a disclosure contained a positive incremental signal about realized earnings direction on average, one would expect it to move forecasts toward the realized outcome in both directions. Instead, the disclosures span a tone gradient, and their effects on the gap track it. The neutral MD&A excerpts, sampled and truncated, produce no detectable average change: $\hat{\beta}_k$ and $\hat{\nu}_k$ are small and insignificant. The cautionary Forward-Looking Statements and Risk Factors tilt the forecast toward a decrease, regardless of the firm's realized direction. These disclosures barely move EPS-decreasing firms: $\hat{\beta}_k$ is +4.5 and +2.2 pp, both insignificant. But for the Forward-

[14] Each disclosure excerpt enters the prompt within a 5,000-token budget, so the longest filings are front-truncated. The estimates below are robust to this truncation: a truncation indicator is insignificant, and the cautionary effects are if anything larger on the non-truncated subsample (Appendix C.2).

Looking Statements and Risk Factors, the interaction $\hat{\nu}_k$ is significant at the 10% and 5% levels respectively, and leaves the EPS-increasing forecasts 23.5 and 25.0 pp further from the outcome than the decreasing cases. The pattern is robust to firm controls, disclosure-by-exposure interactions, and sector heterogeneity (columns (2)–(4)), which makes a training-exposure explanation less likely.

The neutral excerpts add firm content but generate no detectable average response. The cautionary sections widen the gap in a direction consistent with their tone, suggesting that the model continues the tone as a directional tilt rather than extracting only earnings-relevant content. The same firm-year therefore receives a different forecast depending on which disclosure section or excerpt is supplied.

5.2 Supplying the Human's Probability and False Anchors

We next supply the closest information we have to the human belief: the previous-day probability, the latest observed value of the human benchmark for the same forecasting target. This is the prediction-market counterpart of the 10-K disclosures supplied in the analyst setting: the previous-day price is the information native to the market, the signal a participant would actually read. The condition enters the prompt as one added line after the question, stating that as of the previous date the market price for this contract is the given value, "reflecting market participants' current probability estimate"; the number appears as factual context, and the forecast instruction is unchanged (verbatim prompts in Appendix A). In prediction markets these experiments use the minimal single elicitation rather than the persona crowd.[15] To measure how far this condition closes the gap, we stack the minimal

[15]Persona aggregation does not close the gap or reverse its sign (Section 3.1). We use the minimal single elicitation for the remaining prediction-market experiments to isolate the continuation probability and its

Table 6: Signed LLM–Analyst Gap under Context Provision

	(1) Tone	(2) + Firm controls	(3) + Context × Exposure	(4) + Sector het.
Disclosure effect on decreasing firms				
Sampled MD&A	−0.074	−0.070	−0.044	−0.044
	(0.069)	(0.070)	(0.080)	(0.080)
Truncated MD&A	−0.054	−0.048	−0.021	−0.021
	(0.065)	(0.065)	(0.074)	(0.074)
Forward-Looking	+0.030	+0.034	+0.045	+0.045
	(0.068)	(0.069)	(0.080)	(0.080)
Risk Factors	+0.031	+0.035	+0.022	+0.022
	(0.049)	(0.049)	(0.066)	(0.066)
Disclosure × Increase				
× Sampled MD&A	+0.068	+0.063	+0.062	+0.062
	(0.119)	(0.120)	(0.118)	(0.118)
× Truncated MD&A	+0.034	+0.026	+0.025	+0.025
	(0.111)	(0.113)	(0.110)	(0.110)
× Forward-Looking	−0.232*	−0.235*	−0.235*	−0.235*
	(0.120)	(0.121)	(0.120)	(0.120)
× Risk Factors	−0.246**	−0.251***	−0.250**	−0.250**
	(0.096)	(0.096)	(0.098)	(0.098)
Exposure	+0.083	+0.083	+0.098**	+0.179***
	(0.057)	(0.058)	(0.046)	(0.063)
Increase	−0.113	−0.110	−0.110	−0.123
	(0.093)	(0.093)	(0.092)	(0.092)
Exposure × Increase	−0.068	−0.069	−0.069	−0.057
	(0.076)	(0.078)	(0.078)	(0.072)
Firm controls & days-to-resolution FE	No	Yes	Yes	Yes
Context × Exposure	No	No	Yes	Yes
Sector het. (Z, $d_i \times Z$)	No	No	No	Yes
Observations	673,174	653,933	653,933	653,933
R^2	0.111	0.115	0.116	0.123

Notes: OLS of the signed analyst gap on four disclosure-condition indicators and their interaction with the realized outcome, stacked relative to the Yearly+Quarterly EPS baseline, with the exposure and directional structure absorbed throughout. The disclosure effect is β_k on EPS-decreasing firms and $\beta_k + \nu_k$ on EPS-increasing firms. Columns add firm controls and days-to-resolution (weekly) fixed effects (2), the disclosure × exposure interactions (3), and GICS-sector heterogeneity (Z_{it} and $d_i \times Z_{it}$) (4), which leave $\hat{\nu}_k$ essentially unchanged. Sample: S&P 500 firm-years with at least five active analysts; standard errors two-way clustered by company and fiscal year. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

elicitation per event-date and the same prompt with the previous-day probability added, and estimate:

$$\begin{aligned} s_{it,b} = \alpha + \beta\,\mathcal{B}_b + \mu\,(d_i \times \mathcal{B}_b) + \nu\,(g_i \times \mathcal{B}_b) \\ + \lambda d_i + \gamma g_i + \theta\,(d_i \times g_i) + Z_{it}'\chi + (d_i \times Z_{it})'\pi + \Gamma_{it}'\eta + \varepsilon_{it,b}, \end{aligned} \tag{7}$$

where b indexes the two stacked elicitations per event-date and $\mathcal{B}_b = 1$ if elicitation b includes the previous-day probability. The interactions μ and ν let the condition's effect differ by training-corpus exposure and by realized direction, and $\hat{\beta}$ is the condition's effect for unexposed, unfavorable-outcome markets in the interacted specifications; the corresponding effects for other cells add the relevant interactions.

Supplying the previous-day market price closes about half the gap, from −31.3 to −14.5 pp (Table 7), by far the largest closure of any input we supply. The closure is somewhat smaller on markets the model was already exposed to in training (17.5 pp unexposed against 11.8 pp exposed, column (3)), and it varies modestly by outcome direction (−2.1 pp, significant at the 10% level, column (4)). The smaller closure on exposed markets is partly mechanical rather than a difference in the model: those markets already start closer to the human benchmark, so less gap remains for the supplied probability to close. Even so, the residual gap still carries the directional and exposure structure: the outcome term and its interaction with exposure remain large and precisely estimated (+0.42 and +0.12, column (4)). Supplying the human's own probability does not remove it.

This closure does not say how the supplied number enters the forecast: reasoning from the number as information and copying the number produce the same forecast when the

response to prompt form; the resulting treatment effects should be interpreted for that elicitation.

Table 7: Effect of Providing the Previous-Day Human Probability on the Signed Gap

	(1) Core	(2) +Controls	(3) +Cond×Exp	(4) +Cond×Out
Baseline mean $s = -0.313$; with-context mean $s = -0.145$.				
Previous-Day Price	$+0.168^{***}$	$+0.170^{***}$	$+0.175^{***}$	$+0.179^{***}$
	(0.004)	(0.004)	(0.005)	(0.005)
Exposure	-0.016	$+0.011$	$+0.039^{*}$	$+0.038^{*}$
	(0.019)	(0.020)	(0.022)	(0.022)
Outcome (YES)	$+0.403^{***}$	$+0.406^{***}$	$+0.406^{***}$	$+0.417^{***}$
	(0.009)	(0.009)	(0.009)	(0.013)
Exposure × Outcome	$+0.127^{***}$	$+0.124^{***}$	$+0.124^{***}$	$+0.124^{***}$
	(0.023)	(0.023)	(0.023)	(0.023)
Previous-Day Price × Exposure			-0.057^{***}	-0.055^{***}
			(0.012)	(0.013)
Previous-Day Price × Outcome				-0.021^{*}
				(0.011)
Topic (Z, $d \times Z$)	Yes	Yes	Yes	Yes
Source FE, log volume, log duration, weeks-to-resolution FE	No	Yes	Yes	Yes
R^2	0.282	0.287	0.288	0.288
Observations	317,628	316,800	316,800	316,800
Markets	8,131	8,097	8,097	8,097

Notes: Stacked OLS of the outcome-aligned signed gap on the previous-day-probability condition indicator, with the exposure–directional core and heterogeneity (Z_{it}, $d_i \times Z_{it}$; market topic) in every column. The condition extends the minimal single elicitation rather than the informed baseline, so the panel-header baseline mean differs from the informed-baseline gap of Section 3. Columns (2)–(4) add general controls, the previous-day-price × exposure interaction, and the previous-day-price × outcome interaction (the Previous-Day Price row is then the unfavorable-outcome effect). Observations and market counts are reported per column, since the controls are missing on a small number of market-dates. SEs two-way clustered by market and date. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

number is true. Separating the two requires values whose numeric content and information content come apart, so we re-elicit each forecast with the previous-day line filled by one of four values. The *true* value is the reference, the condition behind the closure result above. A *random* value, a real probability drawn from an unrelated event, looks like a market price but carries no information about this event; a model that evaluates what it is given should discount it. The *complement*, $1 - p$, points away from the outcome exactly as strongly as the true value points toward it, the sharpest possible conflict between the number and the event. Each of these three is presented to the model as the correct previous-day probability. A fourth condition, the *disclosed complement*, supplies the same $1 - p$ but tells the model it

is the complement rather than the true value; it separates failing to question a number from failing to process one.

Pooling the three *undisclosed* anchor conditions, indexed by c, we add the supplied value $w_{it,c}$, outcome-aligned as the gap is, to the full specification of Equations (2) and (3):

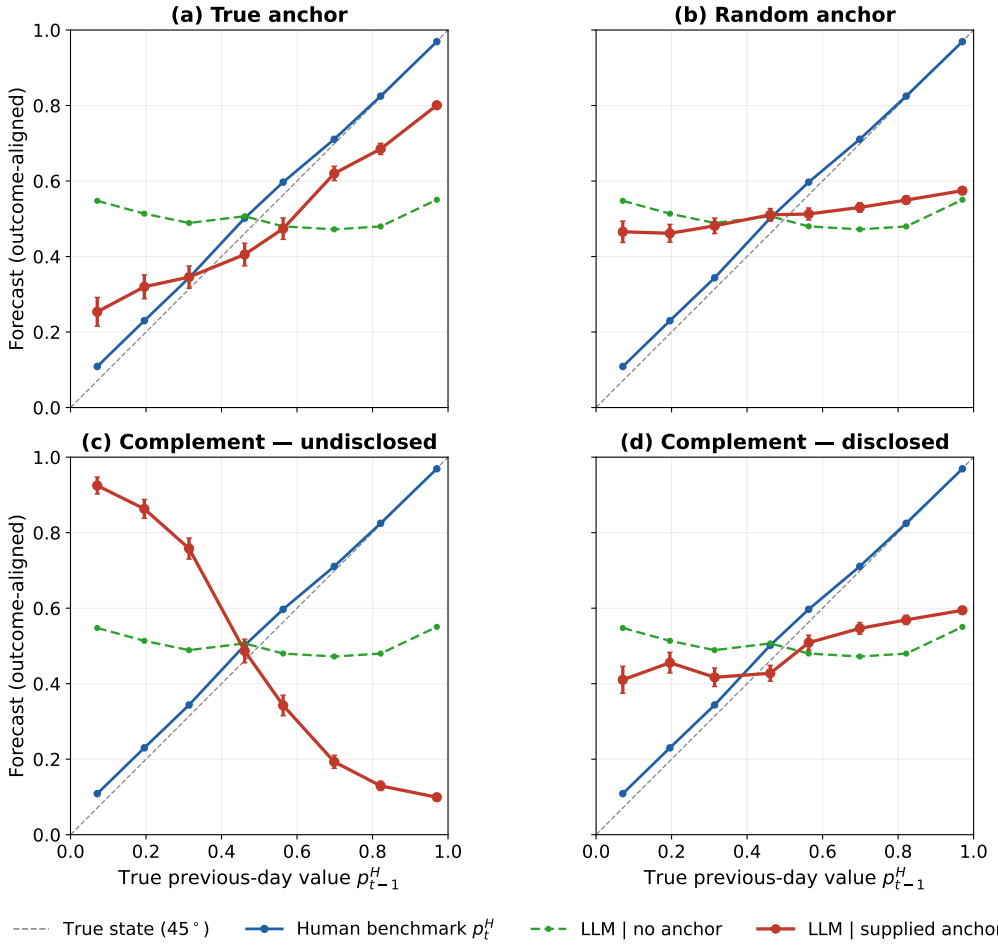
$$\begin{aligned} s_{it,c} = \alpha + \beta\, w_{it,c} + \mu\,(d_i \times w_{it,c}) + \nu\,(g_i \times w_{it,c}) + \phi' \mathbf{a}_c \\ + \lambda d_i + \gamma g_i + \theta\,(d_i \times g_i) + Z'_{it}\chi + (d_i \times Z_{it})'\pi + \Gamma'_{it}\eta + \varepsilon_{it,c}, \end{aligned} \tag{8}$$

where $\mathbf{a}_c$ are fixed effects for the anchor conditions, with the true-value condition as the omitted reference, and μ and ν interact $w_{it,c}$ with exposure and direction. The coefficient β is how much of $w_{it,c}$ the forecast moves with for the same event; because two of the three values are wrong, moving with the number is not reasoning to a correct probability.

The slope on the supplied number is large and robust to controls (Figure 8 and Table 8). Without controls, $\hat{\beta}$ is 0.69 (column (1)). It rises to 0.84 once the exposure and directional core, market controls, and topic fixed effects enter (column (2)), and holds at 0.85 as the supplied-number interactions with exposure and outcome (column (3)) and days-to-resolution fixed effects (column (4)) are added. Its stability across these controls means the anchoring is not an artifact of the event's exposure, direction, or topic. The three anchor conditions share this single slope and differ only in the value placed in the anchor slot; the forecast follows each value, so it tracks the outcome only when the value does. The true value is the human's actual prior-day price, which rises toward the outcome as resolution nears, so the forecast follows it and tracks the human benchmark. The random value carries no real signal, so the forecast is flat. The complement, $1 - p$, is the mirror of the true value,

Figure 8: Mean Forecast Under Genuine and False Anchors

Alt text: Four panels in two rows, one for each supplied-number condition. The true previous-day value runs left to right and the outcome-aligned forecast runs bottom to top, with a dashed diagonal for the true state. A blue line traces the human benchmark and a flat green dashed line the model given no number. A red line with error bars traces the model given the number: it rises with the true value, stays flat when the value is random, and falls steadily from high to low when the complement is supplied without disclosure.



Notes: Mean LLM forecast against the true previous-day market price in prediction markets, under four supplied anchors, with the human benchmark, the no-anchor baseline, and the 45° agreement line shown for reference. Given the genuine previous-day price (a), the forecast rises with the event; given a random number (b), it is flat and matches the no-anchor baseline. Given the complement without disclosure (c), it follows the false number to the wrong side; when told that the value is the complement (d), it inverts it. Bars are 95% confidence intervals for the binned mean, clustered on the market, since a market contributes up to thirty-one dates to the same bin.

Table 8: Forecast Response to the Supplied Number: Specification Build-Up

	(1)	(2)	(3)	(4)
Supplied number	+0.688***	+0.837***	+0.852***	+0.852***
	(0.005)	(0.005)	(0.006)	(0.006)
Exposure		+0.031**	+0.059***	+0.058***
		(0.015)	(0.016)	(0.016)
Outcome		+0.593***	+0.614***	+0.614***
		(0.013)	(0.017)	(0.017)
Exposure × Outcome		+0.035	+0.028	+0.028
		(0.025)	(0.025)	(0.025)
Supplied number × Exposure			−0.048***	−0.048***
			(0.016)	(0.016)
Supplied number × Outcome			−0.043***	−0.043***
			(0.014)	(0.014)
Random anchor (vs. true)		−0.003*	−0.005***	−0.005***
		(0.002)	(0.002)	(0.002)
Complement anchor (vs. true)		+0.049***	+0.055***	+0.055***
		(0.003)	(0.003)	(0.003)
Volume, duration	No	Yes	Yes	Yes
Topic FE + Exposure × Topic	No	Yes	Yes	Yes
Days-to-resolution FE	No	No	No	Yes
Observations	450,930	449,781	449,781	449,781
Clusters (markets)	8,040	8,009	8,009	8,009
R^2	0.359	0.634	0.634	0.636

Notes: Dependent variable: the outcome-aligned signed gap, pooled across three undisclosed anchor conditions (true, random, or complement $1 - p$ previous-day value). In column (1), the supplied-number coefficient is one if the forecast copies the supplied value and zero if it ignores it. In the interacted columns, this coefficient applies to the unexposed, unfavorable-outcome reference cell. The random- and complement-anchor rows are residual level differences relative to the true-anchor condition after controlling for the supplied value. Columns (2)–(4) add the exposure–directional core, controls, and interactions. The sample is the canonical prediction-market panel across both platforms (Polymarket and Kalshi; 30 days before resolution, December-2023 resolutions excluded) restricted to contracts with an elicited anchor. SEs two-way clustered by market and evaluation date. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

so the forecast follows it away from the realized outcome. In each the forecast responds to the number itself, not to which condition supplies it: beyond what their values dictate, the random and complement conditions shift the forecast by only 0.5 and 5.5 pp relative to the true condition (Table 8). The small residual condition effects provide little evidence that the model discounts the two incorrect anchors. That the random-anchor forecast settles at the level of the minimal elicitation, which supplies no number, is consistent with a shared fallback: given no usable signal, supplied or not, the model reverts to a prior of its own, not to the outcome base rate. Told the supplied number is the complement, the model inverts

it, and the forecast rises with the event rather than falling as it does when the complement is undisclosed.

The anchoring barely weakens even where the model might rely on itself. The slope falls by 0.048 on exposed events, where the model has absorbed the outcome from training and so carries its own signal, and by 0.043 on favorable outcomes (Supplied number $\times$ Exposure and $\times$ Outcome, column (3)); both are small against the 0.85 slope. Even there, the anchor overrides the model's own information rather than merely filling an information gap. The exposure and directional terms retained in these specifications show the anchoring adds to the structure of Section 4 rather than replacing it. The halved gap is therefore not a halved belief gap: it measures the model's compliance with the supplied value, and any elicitation that supplies benchmark-tracking numbers can manufacture a small mean gap the same way. The disclosed complement sharpens what governs the forecast: the same number, described differently, produces the opposite forecast, so the model follows the prompt's description of the number rather than the number's relation to the event. The elicitation supplies no verification layer: a pipeline that feeds it a wrong value labeled as true, or a right value labeled wrongly, receives the forecast those labels dictate.

5.3 Framing: Label Order and Wording

Reordering and rewording response labels are equivalent representations of the same event, information, and forecast date. Any resulting movement in the probability is therefore prompt-form variation rather than event variation. Human respondents are themselves sensitive to question form: survey methodology has documented for decades that wording and option order move stated answers (Schuman and Presser, 1981; Schwarz, 1999; Krosnick,

1991). That sensitivity does not carry over to our benchmarks: traders and analysts never see our prompt, and within an event-date the human probability is a single number formed without any researcher's instrument. Framing effects in the elicited probability therefore cannot be read as fidelity to human framing behavior; they are measurement variation in the proxy itself. Reproducing human framing sensitivity could be a legitimate target in survey-simulation applications, but the benchmark would then have to be framed human responses, which our settings do not supply. We examine this by varying the framing in both settings, decomposing the effect into a *positional channel* (the order of the response labels, which carries no event information) and a *connotation channel* (their wording, which introduces training-corpus associations). We use three wording variants: earnings-specific words (*Increase*/*Decrease*, analyst only), plain language (*Yes*/*No*), and semantically weaker labels (A/B). We estimate the framing effects on the minimal elicitation in prediction markets and the EPS-informed prompt in analyst earnings.

Reordering the labels can reverse the LLM's directional call on a material share of events: the flip rate, the share of events on which reordering crosses 0.5, is 11% (Yes/No) and 10% (A/B) in prediction markets, and 17% and 18% under the analyst word labels (Yes/No and Increase/Decrease) versus 5% under the neutral letters (Appendix Figure IA2).[16] The flip is a coarse binary measure; because the model returns a probability, we can measure how far each channel moves the gap and whether it carries a direction. We stack the framing

[16]The flip rates are computed after removing each forecast's estimated exposure component (the fitted $\hat{\lambda}\, d_i + \hat{\theta}\,(d_i \times g_i)$ of Section 4.2), so training-corpus exposure does not explain the flips, though event-specific training memory is not ruled out.

conditions within a setting and estimate

$$
\begin{aligned}
s_{it,c} = \alpha + \lambda\, d_i + \gamma\, g_i + \theta\,(d_i \times g_i) + Z_{it}'\chi + \Gamma_{it}'\eta + (d_i \times Z_{it})'\pi \\
+ \textstyle\sum_{c \neq \text{ref}} \delta_c\, \mathbf{1}[c] + \sum_{c \neq \text{ref}} \tau_c\,(g_i \times \mathbf{1}[c]) + \varepsilon_{it,c},
\end{aligned}
\tag{9}
$$

where $s_{it,c}$ is the outcome-aligned gap for event-date it under framing c, with Yes/No the reference framing in prediction markets and Increase/Decrease in analyst earnings. Each non-reference framing c carries two coefficients relative to it: a level shift δ_c and an outcome interaction τ_c. Because we vary framing within event, the framing coefficients are identified from within-event changes rather than from between-event composition in exposure, direction, or controls.

Across both settings and all specifications, wording changes the gap. The Yes-versus-No and Increase-versus-Decrease labels produce similar patterns ($\hat{\tau}_{\text{Yes/No}} = -0.09$ in analyst earnings, column (4)), while the neutral letters A/B widen the gap on favorable events and narrow it on unfavorable ones ($\hat{\tau}_{A/B} = -0.94$ in prediction markets, -0.69 in analyst earnings, column (4)). When the favorable answer is labeled with a word, Yes or Increase, rather than with the neutral letter A, the model places more probability on it: 44 pp more in prediction markets and 25 pp more in analyst earnings on unfavorable-outcome events (column (4)). The words likely act through their associations: A and B have fewer semantic ties to favorable outcomes, though order and test-answer associations may remain. The connotation channel is thus one source of the documented directional prior.

Reversing the order of the labels operates through the positional channel. Under the word labels, putting the unfavorable label first narrows the gap by 5.5 pp on NO-resolving

Table 9: Framing Effects on the Signed Gap: Wording and Order Channels

	(1) Baseline	(2) +Controls	(3) +Het & Interaction	(4) +Event-time
Panel A: Prediction markets (reference framing: Yes/No)				
No/Yes: same labels, reversed order δ_c	$+0.030^{***}$	$+0.030^{***}$	$+0.055^{***}$	$+0.055^{***}$
	(0.002)	(0.002)	(0.002)	(0.002)
A/B: neutral letters (A = favorable) δ_c	$+0.241^{***}$	$+0.241^{***}$	$+0.441^{***}$	$+0.441^{***}$
	(0.007)	(0.007)	(0.006)	(0.006)
B/A: neutral letters, reversed order δ_c	$+0.218^{***}$	$+0.218^{***}$	$+0.380^{***}$	$+0.380^{***}$
	(0.006)	(0.006)	(0.006)	(0.006)
No/Yes × Outcome τ_c			-0.117^{***}	-0.117^{***}
			(0.004)	(0.004)
A/B × Outcome τ_c			-0.939^{***}	-0.939^{***}
			(0.012)	(0.012)
B/A × Outcome τ_c			-0.760^{***}	-0.760^{***}
			(0.012)	(0.012)
Exposure and outcome terms (d_i, g_i, $d_i \times g_i$)	Yes	Yes	Yes	Yes
Source FE, log volume	No	Yes	Yes	Yes
Heterogeneity (topic, $d\times$ topic)	No	No	Yes	Yes
Framing × outcome	No	No	Yes	Yes
Weeks-to-resolution FE	No	No	No	Yes
Observations	635,256	633,640	633,640	633,640
Clusters (markets)	8,131	8,098	8,098	8,098
Panel B: Analyst earnings (reference framing: Increase/Decrease)				
Dec/Inc: same labels, reversed order δ_c	-0.054^{***}	-0.054^{***}	$+0.047$	$+0.047$
	(0.014)	(0.014)	(0.046)	(0.046)
Yes/No: plain words (Yes = increase) δ_c	-0.052^{***}	-0.052^{***}	$+0.007$	$+0.007$
	(0.012)	(0.012)	(0.009)	(0.009)
No/Yes: plain words, reversed order δ_c	-0.026^{***}	-0.026^{***}	-0.044^{**}	-0.044^{**}
	(0.009)	(0.009)	(0.018)	(0.018)
A/B: neutral letters (A = increase) δ_c	-0.203^{***}	-0.203^{***}	$+0.248^{***}$	$+0.248^{***}$
	(0.054)	(0.054)	(0.091)	(0.091)
B/A: neutral letters, reversed order δ_c	-0.159^{***}	-0.159^{***}	$+0.112$	$+0.112$
	(0.036)	(0.036)	(0.084)	(0.084)
Dec/Inc × Outcome τ_c			-0.155^{**}	-0.155^{**}
			(0.075)	(0.075)
Yes/No × Outcome τ_c			-0.090^{***}	-0.090^{***}
			(0.022)	(0.022)
No/Yes × Outcome τ_c			$+0.028$	$+0.028$
			(0.035)	(0.035)
A/B × Outcome τ_c			-0.694^{***}	-0.694^{***}
			(0.141)	(0.141)
B/A × Outcome τ_c			-0.416^{***}	-0.416^{***}
			(0.137)	(0.137)
Exposure and outcome terms (d_i, g_i, $d_i \times g_i$)	Yes	Yes	Yes	Yes
Firm controls	No	Yes	Yes	Yes
Heterogeneity (sector, $d\times$ sector)	No	No	Yes	Yes
Framing × outcome	No	No	Yes	Yes
Weeks-to-resolution FE	No	No	No	Yes
Observations	795,384	795,384	795,384	795,384
Clusters (firms)	578	578	578	578

Notes: Stacked regression per panel on the baseline elicitation; each framing of an event-date is one row, with the reference framing (panel header) omitted and *reversed order* listing the unfavorable label first. Each framing carries a level shift δ_c and an outcome interaction τ_c (column (3)); column (4) adds event-time fixed effects (weekly bins in both settings). The exposure and control specification follows Table 4; the displayed exposure and directional coefficients are estimated on the framing samples. Standard errors two-way clustered (market × date; firm × year); the second dimension is constant across columns at 1,382 evaluation dates in Panel A and five fiscal years in Panel B. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

prediction markets and an insignificant 4.7 pp on EPS-decreasing firm-years, but widens it by 6.2 pp on YES-resolving markets and 10.8 pp on EPS-increasing firm-years (Table 9, column (4)). The neutral letters produce larger order effects. Putting B before A widens the gap by 6.1 pp on NO-resolving markets and 13.6 pp on EPS-decreasing firm-years, but narrows it by 11.8 pp on YES-resolving markets and 14.2 pp on EPS-increasing firm-years. With no meaning to anchor on, the model favors the label listed first. The order effect, however, has no consistent direction: its sign varies with the framing, making it a sensitivity to check rather than a correctable bias. In either channel, the retained exposure and directional terms show that label-form effects add to the structure documented earlier rather than replace it.

5.4 Reasoning: Instruction and Chain-of-Thought

Instructing the model to deliberate before answering improves its performance on reasoning tasks (Wei et al., 2022; Kojima et al., 2022), so one might expect it to move the elicited probability toward the human benchmark as well. We add two single-elicitation reasoning conditions: *Reasoning Instruction*, which tells the model to think before returning a single-token answer, and *Chain-of-Thought (CoT)*, which asks it to write explicit reasoning first, implemented in two passes: the model first writes its reasoning under an instruction not to state a final Yes-or-No answer, and the trace is then supplied to the forecast prompt, so the answer token responds to the reasoning itself rather than to an answer already written in it (verbatim prompts in Appendix A). The Reasoning Instruction condition generates no intermediate reasoning tokens, whereas CoT supplies a visible generated trace. Their contrast therefore measures the incremental effect of adding that trace, conditional on the two

prompt designs. Reasoning changes the instruction, not the information, so any movement in the gap is an instruction effect. We stack the baseline and the two reasoning conditions, one row per event-date per condition, and estimate:

$$\begin{aligned} s_{it,r} = \alpha + \lambda\, d_i + \gamma\, g_i + \theta\,(d_i \times g_i) + Z_{it}'\chi + (d_i \times Z_{it})'\pi + \Gamma_{it}'\eta \\ + \delta_{\text{instr}}\mathbf{1}[r{=}\text{instr}] + \delta_{\text{CoT}}\mathbf{1}[r{=}\text{CoT}] \\ + \rho_{\text{instr}}\,(g_i \times \mathbf{1}[r{=}\text{instr}]) + \rho_{\text{CoT}}\,(g_i \times \mathbf{1}[r{=}\text{CoT}]) + \varepsilon_{it,r}, \end{aligned} \tag{10}$$

where δ_r is condition r's level effect on the gap relative to the no-reasoning reference, ρ_r is its outcome interaction, and the remaining terms are as in Equation 3.[17]

Reasoning does not close the gap (Table 10): in prediction markets the reasoning instruction narrows it by only 1.8 pp and chain-of-thought by 1.1 pp ($\hat{\delta}_{\text{instr}}$ and $\hat{\delta}_{\text{CoT}}$, closing 6% and 4% of the baseline gap), while in analyst earnings chain-of-thought leaves it unchanged and the reasoning instruction *widens* it by 7.8 pp, from −16.4 to −24.2 pp. Reasoning also does not remove the directional structure. In the fully interacted specifications, the reasoning instruction narrows the gap on the unfavorable direction but widens it on the favorable one ($\hat{\rho}_{\text{instr}} < 0$ in both settings); its pooled analyst effect nevertheless widens the average gap. It lowers the stated favorable probability regardless of the realized outcome rather than extracting evidence.

[17] The estimation sample is the training-corpus exposure panel restricted to market-dates with all three reasoning conditions available: the reasoning instruction has full coverage, while chain-of-thought terminates in a parseable answer on 91% of the panel's market-dates (99% on Kalshi and 91% on Polymarket), leaving 145,115 of the 158,814. In analyst earnings all three conditions have full coverage on the canonical panel.

Table 10: Reasoning Effect and Directional Asymmetry: Combined Reasoning Conditions

	(1)	(2)	(3)	(4)
Panel A: Prediction Markets				
Exposure	−0.001	+0.022	+0.019	+0.018
	(0.018)	(0.018)	(0.024)	(0.024)
Outcome (YES)	+0.393***	+0.404***	+0.404***	+0.430***
	(0.013)	(0.013)	(0.013)	(0.014)
Exposure × Outcome	+0.139***	+0.136***	+0.130***	+0.131***
	(0.030)	(0.030)	(0.030)	(0.030)
Reasoning Instruction	+0.018***	+0.018***	+0.018***	+0.035***
	(0.002)	(0.002)	(0.002)	(0.002)
Chain-of-thought	+0.011**	+0.011**	+0.011**	+0.011*
	(0.004)	(0.004)	(0.004)	(0.006)
Reasoning Instruction × Outcome				−0.079***
				(0.004)
Chain-of-thought × Outcome				+0.002
				(0.009)
Source FE, Log Volume	No	Yes	Yes	Yes
Exposure × Topic interactions	No	No	Yes	Yes
Calendar controls	No	No	No	Yes
Observations	435,345	434,706	434,706	434,706
Markets	7,364	7,340	7,340	7,340
Panel B: Analyst Earnings				
Exposure	+0.091**	+0.095**	+0.248***	+0.247***
	(0.041)	(0.043)	(0.059)	(0.055)
Outcome (Increase)	−0.309***	−0.313***	−0.332***	−0.153
	(0.022)	(0.022)	(0.030)	(0.093)
Exposure × Outcome	−0.027	−0.032	−0.014	−0.013
	(0.099)	(0.100)	(0.098)	(0.097)
Reasoning Instruction	−0.078***	−0.078***	−0.078***	+0.145**
	(0.021)	(0.021)	(0.021)	(0.068)
Chain-of-thought	+0.001	+0.001	+0.001	+0.127
	(0.021)	(0.021)	(0.021)	(0.090)
Reasoning Instruction × Outcome				−0.343***
				(0.104)
Chain-of-thought × Outcome				−0.193
				(0.153)
Firm controls (log analysts, log assets, ROA, B/M, cash)	No	Yes	Yes	Yes
Exposure × GICS Sector interactions	No	No	Yes	Yes
Calendar controls	No	No	No	Yes
Observations	397,740	397,740	397,740	397,740
Firms	578	578	578	578

Notes: Dependent variable: the outcome-aligned signed gap. Panel A is the prediction-market lookahead panel, Panel B the canonical analyst panel; both stack reasoning-instruction and chain-of-thought forecasts onto the no-reasoning baseline. Columns (2)–(4) add controls, exposure-by-category interactions, and reasoning × outcome interactions, so the reasoning main effects are the unfavorable-outcome effect. SEs two-way clustered by market/date (A), company/year (B). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

6 Downstream Consequences for Financial Inference

A researcher who lacks a survey of human forecasts may generate the missing expectation with a language model, and machine forecasts are already benchmarked against analysts' (van Binsbergen et al., 2023; Cao et al., 2024). We ask what that substitution does in two regressions economists use. The first is an announcement-window return validation motivated by earnings-response designs (Ball and Brown, 1968; Collins and Kothari, 1989); because our annual directional outcome is not the quarterly earnings surprise, we do not interpret it as a conventional earnings-response coefficient. The second is the belief-rigidity regression, in which predictable forecast errors reveal how beliefs update (Coibion and Gorodnichenko, 2015; Bouchaud et al., 2019). The analyst consensus and the elicited probability are two forecasts of the same annual direction, so a valid stand-in should preserve the consensus's empirical association in each design. The corrections of Section 4.3 provide a direct test of whether removing the components documented earlier is enough to do so.

6.1 Announcement Returns

The announcement-window exercise is run on a directional *forecast residual*: the outcome minus what the forecaster expected. We measure that expectation as a probability, so $U_i^f = g_i - \tilde{p}_i^f$ is the realized direction of the fiscal-year EPS change, net of the probability forecaster f assigned to an increase, and β_f measures how strongly announcement returns respond to it.[18] The object is useful as a comparative validation because the human and

[18] The classic design scales a point-forecast error by price; the probability shortfall $g_i - \tilde{p}_i^f$ is not that object. Its closest precedent is Veenman and Verwijmeren (2018), who price an ex ante probability of analyst-forecast pessimism built from analysts' past forecast errors.

machine series enter the same regression, but it is not a canonical quarterly earnings surprise.

For each quarterly earnings announcement we take each forecaster's probability from the last elicitation preceding it, discard matches staler than 45 days, and estimate

$$CAR_i[-1,+1] \;=\; \alpha_f + \beta_f\, U_i^f + \omega_1 g_i + \omega_2 (d_i \times g_i) + X_i'\zeta + \psi_{q(i)} + \varepsilon_i \tag{11}$$

separately for four forecasters, with one observation per announcement: 8,282 announcements over 567 firms, at a median lag of seven days between the elicitation and the announcement.[19] Each elicitation is a forecast about the fiscal year, while each observation is one quarterly announcement inside that year. For the first three quarters the fiscal-year outcome g_i is therefore not yet known when the announcement occurs, so U_i^f carries information that arrives after the event. Holding g_i fixed removes it. From column (2) on, the fiscal-year outcome and its interaction with training-corpus exposure are controls, so the specification conditions on both outcome direction and the outcome-dependent exposure term used by the ex post correction. The comparison runs across announcements that ended the year the same way, and β_f is identified from the forecaster's probability alone. Among those announcements, β_f measures the conditional association between the pre-announcement probability and the announcement return; it does not identify the market response to a contemporaneous quarterly earnings surprise. All four forecasters are treated identically, so the comparison across them is unaffected, and we read β_f on that basis as a conditional return association

[19] $CAR_i[-1,+1]$ cumulates firm i's return net of the value-weighted market from the trading day before the announcement to the trading day after it; the announcement window is where the price response to earnings news concentrates (Beaver, 1968). Announcement dates and realized quarterly EPS are from I/B/E/S, daily returns from CRSP. X_i holds log market capitalization, the prior return, prior volatility, analyst coverage, and the scaled magnitude of the realized earnings change; $\psi_{q(i)}$ are calendar year–quarter effects of the quarterly period end (27 in the estimation sample), which absorb training-corpus exposure because exposure is a function of the fiscal year. Standard errors are clustered on the firm.

rather than as a contemporaneous earnings-response coefficient.

The four forecasters are the analyst consensus, the raw LLM probability, and its ex ante and ex post linear corrected scores from Section 4.3, each formed by subtracting the held-out fitted gap of Equation (5). The ex ante correction uses only what a researcher observes when the probability is elicited, namely training-corpus exposure, sector, firm fundamentals, and the forecast horizon; the ex post correction adds the realized outcome and its interaction with exposure, so it bounds what the decomposition accounts for rather than supplying an adjustment anyone could form in advance. Both re-center the elicitation on the consensus at the matched announcements, closing the average *raw-scale* gap from −11.7 percentage points to +0.4 ex ante and +0.5 ex post, on firms the estimation never saw. The linear corrected scores enter U_i^f unclipped, as Section 4.3 constructs them; bounded variants lead to the same conclusion.[20] If the elicited probability is a valid stand-in for the consensus, β_{LLM} should match β_H; if the correction repairs the substitution, the corrected coefficient should sit closer to β_H than β_{LLM} does.

Table 11 reports the estimates, adding controls in the same order within each panel. The analyst consensus is positively associated with announcement returns throughout. Its coefficient moves only from +0.0269 without controls to +0.0247 under the full set, significant at the one percent level in every column, and a one-standard-deviation surprise moves the three-day return by 0.8 percentage points. The association is not subsumed by the quarter's conventional earnings surprise: column (5) adds that variable and leaves the human coefficient at +0.018 ($t = 5.7$).[21] The elicited probability behaves differently. It starts significant

[20] With the corrected scores clipped to $[0, 1]$, the fully controlled coefficients are +0.0026 ($t = 0.7$) ex ante and +0.0014 ($t = 0.4$) ex post.

[21] The conventional surprise is the quarter's reported EPS less the analyst consensus estimate for that quarter, scaled by the pre-announcement price and winsorized at the first and ninety-ninth percentiles.

at +0.0112 and falls as the controls enter, reaching +0.0021 ($t = 0.6$) once the realized direction, quarter effects, and firm characteristics are absorbed. Neither correction changes the pattern: the ex ante and ex post series follow the raw one down to +0.0023 and +0.0017, both indistinguishable from zero, and against the conventional surprise in column (5) all three machine series sit at or below +0.0006 with $t < 0.7$. Nor are the machine coefficients merely imprecise, or small only because the two surprises are scaled differently. A one-standard-deviation machine surprise moves the return by 0.06 to 0.10 percentage points, against 0.8 for the human. A firm-level bootstrap rejects equality with the human coefficient for all three series ($z \geq 5.8$, and 4.6 to 5.8 in per-standard-deviation units). When the human surprise and the ex ante corrected surprise enter the same regression, the human's coefficient holds at +0.025 while the corrected one sits at −0.004 ($t = -1.2$), despite a correlation of 0.55 between them.[22]

The human coefficient is anticipatory, and the split by fiscal quarter shows it. At the interim announcements, where the fiscal-year direction is still in play, the human coefficient is essentially the pooled one, +0.0268 ($t = 8.6$) under the full controls, and every machine series stays at zero ($|t| \leq 1.1$). At the annual announcement, all four surprises are unpriced, the human's included ($|t| < 0.5$): the direction is close to resolved by then, and the most recent elicitation preceding the announcement is a median five weeks old, so the cell carries little news for any forecaster.[23] The comparison the table runs is therefore about the months

[22]The raw and ex post series give −0.0038 ($t = -1.1$) and −0.0045 ($t = -1.3$) in the same encompassing regression. Adding the full correction inputs—sector and exposure × sector effects, annual firm characteristics, and horizon effects—alongside the announcement controls gives +0.0250 (SE 0.0032) for the analyst consensus, +0.0020 (0.0035) for the raw LLM, +0.0015 (0.0035) ex ante, and +0.0016 (0.0034) ex post. Thus the conclusion does not rely on claiming that the displayed return controls absorb every correction regressor.

[23]Of the 8,156 announcements matched to a Compustat fiscal quarter, 7,105 are interim (fiscal quarters one to three, median elicitation lag six days) and 1,051 are annual (fiscal quarter four, median lag 36 days); annual announcements are underrepresented because the elicitation panel thins ahead of them, so few pass

before resolution, which is where a forecast is useful, and that is precisely where the machine series carry none of the consensus's information.

The correction therefore removes the average raw-scale gap without recovering the human return association. Under the full controls the corrected surprise explains 0.007% of return variation against the human's 0.97%, a ratio of roughly 140 to one, and the shortfall is not a matter of scale: rescaling a surprise rescales its coefficient and standard error together, leaving the t-statistic, the partial R^2, and the encompassing comparison unchanged. A correction built from exposure, sector, and firm characteristics removes the level of the elicited probability, not the cross-sectional variation the regression consumes. The average gap closes because the individual gaps come to cancel, not because they shrink: at the typical announcement the ex ante corrected score sits 29 percentage points from the consensus, no closer than the uncorrected one at 28, and the ex post version still sits 26 away.

6.2 Belief Rigidity

The belief-rigidity regression is estimated from belief revisions rather than from a cross-section of surprises. Substituting the LLM probability for the analyst consensus therefore tests whether it reproduces how beliefs change over time, not only their average level, which the earlier sections have already characterized. For each of the same four forecasters — the analyst consensus, the raw LLM probability, and its ex ante and ex post linear corrected scores — the *forecast error* is $\mathrm{fe}^f_{it} \equiv g_i - \tilde{p}^f_{it}$ and the *revision* since the previous update is $\mathrm{rev}^f_{it} \equiv \tilde{p}^f_{it} - \tilde{p}^f_{i,t-1}$, both defined on the raw increase probability $\tilde{p}^f_{it}$ rather than its outcome-

the staleness screen. Within the annual subsample the human–machine differences are also zero ($|z| \leq 0.3$).

Table 11: Announcement Returns on the Directional Forecast Residual

Forecaster f substituted into Eq. (11)	(1) No controls	(2) +Outcome terms	(3) +Quarter FE	(4) +Firm controls	(5) +Conventional surprise
Panel A: Analyst consensus					
Slope β_f	$+0.0269^{***}$	$+0.0255^{***}$	$+0.0260^{***}$	$+0.0247^{***}$	$+0.0183^{***}$
	(0.0026)	(0.0029)	(0.0030)	(0.0031)	(0.0032)
Partial R^2	0.016	0.011	0.011	0.010	0.005
Panel B: LLM probability, uncorrected					
Slope β_f	$+0.0112^{***}$	$+0.0067^{***}$	+0.0052	+0.0021	+0.0006
	(0.0018)	(0.0023)	(0.0035)	(0.0035)	(0.0036)
Partial R^2	0.005	0.001	0.000	0.000	0.000
Panel C: LLM linear corrected score, ex ante					
Slope β_f	$+0.0113^{***}$	$+0.0067^{***}$	$+0.0056^{*}$	+0.0023	+0.0002
	(0.0018)	(0.0023)	(0.0033)	(0.0033)	(0.0033)
Partial R^2	0.005	0.001	0.001	0.000	0.000
Panel D: LLM linear corrected score, ex post					
Slope β_f	$+0.0111^{***}$	$+0.0056^{**}$	+0.0046	+0.0017	+0.0002
	(0.0021)	(0.0024)	(0.0034)	(0.0034)	(0.0035)
Partial R^2	0.004	0.001	0.000	0.000	0.000
Realized direction g_i	No	Yes	Yes	Yes	Yes
Exposure × realized direction	No	Yes	Yes	Yes	Yes
Fiscal-quarter FE	No	No	Yes	Yes	Yes
Firm controls	No	No	No	Yes	Yes
Conventional surprise	No	No	No	No	Yes
Announcements	8,282	8,282	8,282	8,282	8,282

Notes: The coefficient β_f from Equation (11), with $CAR_i[-1,+1]$ as the dependent variable and one observation per quarterly announcement, over 567 firms; announcements listed more than once (identical duplicate records, renamed tickers, dual share classes) enter once, keeping the fresher elicitation. Each panel is a separate regression, and each column within a panel is a separate regression: the surprise $U_i^f = g_i - \tilde{p}_i^f$ is computed from that panel's probability series, read from the last elicitation preceding the announcement, and the four panels are never estimated jointly. Controls enter cumulatively across columns; the conventional surprise in column (5) is the quarter's reported EPS less the analyst consensus estimate, scaled by the pre-announcement price and winsorized at the first and ninety-ninth percentiles. Partial R^2 is the share of return variation the directional surprise explains beyond that column's controls. From column (2) on, both g_i and $d_i \times g_i$ enter. Panels C and D replace the LLM probability with the ex ante and ex post linear corrected scores, each formed by two-fold cross-fitting across firms and entering unclipped as constructed in Section 4.3. Training-corpus exposure is a function of the fiscal year and is absorbed by the quarter effects from column (3) on. Standard errors in parentheses are clustered on the firm. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

aligned transform. Regressing the error on the revision,

$$\mathrm{fe}^{f}_{it} \;=\; a_f + b_f\,\mathrm{rev}^{f}_{it} + \Gamma'_{it}\eta_f + \varepsilon^{f}_{it}, \tag{12}$$

gives the slope b_f, where Γ_{it} collects the controls of Table 12. This is the design of Coibion and Gorodnichenko (2015), carried from point forecasts to the probability scale and, as in analyst applications (Bouchaud et al., 2019), run on a fixed-event annual target rather than a rolling fixed horizon, with the forecast-horizon effects below absorbing the shrinking horizon. Under full-information rational expectations the forecast error is unpredictable from anything the forecaster knows, including the forecaster's own revision, so b_f is zero. A positive b_f indicates underreaction; a negative b_f is read as overreaction. The sign of b_f alone thus decides the inference, so a sign change reverses the conclusion a researcher draws from the same data.

Estimated on the analyst firm-year–date panel, with standard errors two-way clustered by firm and fiscal year, Equation (12) gives the standard underreaction coefficient for the consensus, $b_H = +0.68$ ($t = 12.5$), positive and of the same order as consensus-level estimates in point-forecast settings, and the opposite sign for the LLM probability, $b_{\mathrm{LLM}} = -0.43$ ($t = -11.0$).

Table 12 reports the slopes as controls accumulate across columns, from the bare regression to forecast-horizon fixed effects, sector fixed effects, and firm fundamentals with analyst coverage. Across every specification the coefficients are essentially unchanged: the analyst slope falls only from 0.68 to 0.63 and the LLM slope stays near -0.43. The reversal is a property of the panel, not of a specification; and because the regression is run in probability

space, on revisions measured days rather than quarters apart, the magnitude of b_f does not map to a structural rigidity parameter.

The reversal does not come from the gap components documented in earlier sections. Replacing the LLM probability with its linear corrected score from Section 4.3 leaves the slope at -0.41 under the ex ante correction and at -0.46 under the ex post correction, which uses the realized outcome and removes almost all of the level gap (Table 12, Panels C and D).[24] Both are stable as controls accumulate and neither approaches the analyst slope. If removing the structure the paper documents were enough to restore the human sign, it would show here. The ex post correction is the sharper test: it removes the directional prior that an ex ante correction cannot reach, and the slope moves further from the analyst's rather than toward it. Nor is the negativity the overreaction of diagnostic-expectations models, the pattern Bordalo et al. (2020) document for individual forecasters in the same survey panels whose consensus underreacts. A diagnostic forecaster over-responds to *real* information, so its revisions predict the outcome; the LLM's revisions predict the outcome with an R^2 of only 0.001, indicating little incremental outcome signal to over-respond to.

The LLM probability still moves between consecutive dates, and that movement admits two readings. In the first, the elicited probabilities are snapshots of one belief the model updates between dates: the revision reflects arriving news and should co-move with the consensus revision, which responds to the same news. In the second, each elicitation is a fresh continuation of that date's prompt, and no belief carries over between dates: the revision still moves, but it reflects the model's sensitivity to small changes in the prompt

[24] The bounded ex ante log-odds correction of Appendix Table IA9 likewise produces a negative full-control slope, -0.369 (two-way clustered SE 0.041).

Table 12: Forecast Errors on Forecast Revisions

Forecaster f substituted into Eq. (12)	(1) Baseline	(2) +Horizon FE	(3) +Sector FE	(4) +Full Controls
Panel A: Analyst consensus				
Slope b_f	$+0.682^{***}$	$+0.635^{***}$	$+0.621^{***}$	$+0.626^{***}$
	(0.054)	(0.040)	(0.037)	(0.038)
Panel B: LLM probability, uncorrected				
Slope b_f	-0.434^{***}	-0.435^{***}	-0.430^{***}	-0.420^{***}
	(0.040)	(0.044)	(0.044)	(0.042)
Panel C: LLM linear corrected score, ex ante				
Slope b_f	-0.413^{***}	-0.420^{***}	-0.409^{***}	-0.407^{***}
	(0.025)	(0.031)	(0.030)	(0.028)
Panel D: LLM linear corrected score, ex post				
Slope b_f	-0.469^{***}	-0.472^{***}	-0.462^{***}	-0.460^{***}
	(0.015)	(0.017)	(0.017)	(0.015)
Horizon (event-week) FE	No	Yes	Yes	Yes
Sector FE	No	No	Yes	Yes
Firm controls & coverage	No	No	No	Yes
Observations (per forecaster)	129,936	129,936	129,936	129,936

Notes: The coefficient b_f from Equation (12), estimated separately for each forecaster on the same firm-year–date revision panel. $b_f > 0$ indicates underreaction; $b_f < 0$ apparent overreaction. The four panels are four separate regressions of Equation (12), one per forecaster: within a panel the forecast error and the revision are both computed from that panel's probability series, so the entries are not coefficients from a single specification. Controls enter cumulatively across columns: forecast-horizon (event-week) fixed effects, GICS sector fixed effects, and firm fundamentals with log analyst coverage. Panels C and D replace the LLM probability with its ex ante and ex post linear corrected scores, each formed by two-fold cross-fitting across firms. The sample requires the firm controls used to build the correction, giving 129,936 revision observations from 578 companies and 2,644 firm-years. Standard errors (in parentheses) are two-way clustered by firm and fiscal year; the fiscal-year dimension has five clusters. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

and should be unrelated to the consensus revision. Equation (12) cannot separate the two, since it regresses each forecaster's error on its own revision and the two series never enter the same regression. Projecting the LLM revision on the analyst revision,

$$\mathrm{rev}_{it}^{\mathrm{LLM}} \;=\; a_\rho + \rho\,\mathrm{rev}_{it}^{H} \;+\; \Delta u_{it}, \tag{13}$$

divides it into a consensus-tracking part $\rho\,\mathrm{rev}^H$ and a residual Δu, the consensus-orthogonal revision component. The LLM revision has a standard deviation of 0.069 against 0.038 for the consensus revision, but the projection, estimated without controls and on the same sample

as the baseline column, and with an intercept as shown, returns $\hat{\rho} = 0.09$ ($t = 13.8$) and an R^2 of 0.003. The correction of Section 4.3 leaves this residual intact, which is why it left the slope where it found it, and the reason is that its inputs barely vary within a firm-year. Exposure, sector, and firm characteristics are constant across the dates of a given firm-year, and only the horizon profile moves, so the fitted gap subtracted at consecutive dates is almost the same number, and subtracting a constant leaves a difference unchanged. The standard deviation of the correction's change between consecutive dates is 0.015, against 0.069 for the LLM revision itself, and the corrected and raw revisions correlate at 0.98. Substituting Equation (13) into the baseline slope, $b_{\mathrm{LLM}} = \mathrm{Cov}(\mathrm{fe}^{\mathrm{LLM}}, \mathrm{rev}^{\mathrm{LLM}})/\mathrm{Var}(\mathrm{rev}^{\mathrm{LLM}})$, splits it into the two corresponding channels,

$$b_{\mathrm{LLM}} = \underbrace{\frac{\mathrm{Cov}(\mathrm{fe}^{\mathrm{LLM}},\, \rho\, \mathrm{rev}^{H})}{\mathrm{Var}(\mathrm{rev}^{\mathrm{LLM}})}}_{\text{consensus-tracking} = +0.021} + \underbrace{\frac{\mathrm{Cov}(\mathrm{fe}^{\mathrm{LLM}},\, \Delta u)}{\mathrm{Var}(\mathrm{rev}^{\mathrm{LLM}})}}_{\text{consensus-orthogonal} = -0.454} = -0.43. \tag{14}$$

The consensus-tracking channel is small and weakly positive, the same sign as the analyst coefficient. The negative covariance is concentrated in the consensus-orthogonal component. If that component behaves as transitory elicitation error, the sign follows mechanically: an error u_t that pushes the probability up makes the forecast error more negative and the revision more positive, so $\mathrm{Cov}(-u_t,\, u_t - u_{t-1}) < 0$ unless the error is nearly perfectly persistent. The projection does not identify Δu as noise, because it can also contain private model-side information or differences in timing. Its low outcome-predictive content and its stability after correction are nevertheless consistent with re-elicitation variation being the dominant component. The split is similar across Table 12: the consensus-tracking channel is about +0.02 and the consensus-orthogonal channel about −0.45, with corresponding values

of $+0.02$ and -0.44 after correction.

The two applications therefore share a measurement lesson, although their mechanics differ. In the return exercise, the correction changes the mean and the component explained by its observables but leaves little of the human forecast's relevant cross-sectional association; this remains true when the full correction inputs enter the return regression. In the rigidity exercise, most correction inputs are constant within a firm-year and largely difference out, so the corrected revision remains close to the raw revision. Average agreement is consequently neither necessary nor sufficient for agreement in a downstream coefficient. A machine-generated forecast residual can match the human residual's mean while failing to carry its return association, and a corrected level can still generate revisions unlike the human revisions. The substitution must therefore be validated on the economic estimand a study reports, not only on the average probability from which that estimand is constructed.

7 Conclusion

In the two settings we study, an LLM's elicited probability is not a reliable measure of the human belief it is meant to represent. It departs from the human consensus by a large and structured margin that richer prompts do not close, because the number is a continuation of the prompt rather than a stored belief. The elicited probability responds to the form of elicitation as much as to its content, so its components can offset: the average gap can be small while individual forecasts remain far from the human on both sides, or while the model is doing no more than anchoring on a supplied value. Agreement with a human benchmark on average is therefore not sufficient to treat an LLM as a synthetic respondent: validity

has to be established against ground truth in the specific model, prompt, and setting of use, and cannot be carried over from another model. Nor is the measurement error classical noise that merely attenuates coefficients toward zero: its components are structured and can offset, so its effect on a downstream analysis need not reduce to attenuation and depends on the empirical design.

The downstream exercises make that distinction concrete. The linear cross-fitted correction removes the average raw-scale gap without reducing individual mean absolute error; bounded variants reduce that loss, but the corrected series recover neither the analyst consensus's conditional association with announcement returns nor its positive belief-rigidity coefficient. Correction is therefore useful as a diagnostic and, in bounded form, can improve forecast loss, but it does not by itself validate an LLM probability for an economic estimand.

The same structure that makes the raw probability unreliable can be estimated without access to model weights or training data, but the exercise still requires stated probabilities, realized outcomes, cutoff-based exposure status, and a validation sample of human benchmarks. Exposure- and prompt-form components may then inform corrections in settings where their regressors are observed; outcome-conditioned components remain ex post diagnostics. Each of these sensitivities enters at a different stage of an agent pipeline: the text a retrieval step supplies, the number a tool returns, the reasoning the agent runs, or the template that formats its queries. A pipeline that chains these steps is subject to all of them at once; because the errors are systematic, they need not average out across stages and may instead offset or compound. As academic and commercial practice increasingly substitutes LLM-generated populations for human samples, these diagnostics are what separate a measured belief from a completion shaped by the training corpus and the prompt's surface

form.

References

Angelopoulos, A. N., S. Bates, C. Fannjiang, M. I. Jordan, and T. Zrnic (2023). Prediction-powered inference. *Science 382*(6671), 669–674. arXiv:2301.09633.

Ball, R. and P. Brown (1968). An empirical evaluation of accounting income numbers. *Journal of Accounting Research 6*(2), 159–178.

Bastianello, F. (2025). Bias and Predictability in Analysts' Beliefs. SSRN Working Paper 5733302.

Bastianello, F., P. Decaire, and M. Guenzel (2024). Mental Models and Financial Forecasts. SSRN Working Paper 5004839.

Beaver, W. H. (1968). The information content of annual earnings announcements. *Journal of Accounting Research 6*, 67–92. Empirical Research in Accounting: Selected Studies 1968, Supplement.

Bhagwat, V., J. A. Cookson, C. Dim, and M. Niessner (2026). The Market's Mirror: Revealing Investor Disagreement with LLMs. SSRN Working Paper 5375473.

Bordalo, P., N. Gennaioli, Y. Ma, and A. Shleifer (2020). Overreaction in macroeconomic expectations. *American Economic Review 110*(9), 2748–2782.

Bouchaud, J.-P., P. Krüger, A. Landier, and D. Thesmar (2019). Sticky expectations and the profitability anomaly. *Journal of Finance 74*(2), 639–674.

Bursztyn, L., T. Chaney, T. A. Hassan, and A. Rao (2024). The Immigrant Next Door. *American Economic Review 114*(2), 348–384.

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

Cao, S., W. Jiang, J. Wang, and B. Yang (2024). From man vs. machine to man + machine: The art and AI of stock analyses. *Journal of Financial Economics 160*, 103910.

Chen, H., A. Didisheim, and L. Somoza (2026). Out of the black box: Uncertainty quantification for LLMs via conditional probabilities. NBER Working Paper 34965; SSRN Working Paper 6425017.

Christensen, T., S. Gonçalves, and B. Perron (2026). Bootstrapping with AI/ML-generated labels. arXiv:2604.23770.

Coibion, O. and Y. Gorodnichenko (2015). Information rigidity and the expectations formation process: A simple framework and new facts. *American Economic Review 105*(8), 2644–2678.

Collins, D. W. and S. P. Kothari (1989). An analysis of intertemporal and cross-sectional determinants of earnings response coefficients. *Journal of Accounting and Economics 11*(2–3), 143–181.

Diercks, A. M., J. D. Katz, and J. H. Wright (2026). Kalshi and the Rise of Macro Markets. NBER Working Paper 34702.

Dupas, P., A. Handlan, A. S. Modestino, M. Niederle, M. Seré, H. Sheng, J. Wolfers, and

Seminar Dynamics Collective (2026). Gender Differences in Economics Seminars. *American Economic Review 116*(2), 749–789.

Easterwood, J. C. and S. R. Nutt (1999). Inefficiency in analysts' earnings forecasts: Systematic misreaction or systematic optimism? *The Journal of Finance 54*(5), 1777–1797.

Fedyk, A., A. Kakhbod, P. Li, and U. Malmendier (2024). AI and Perception Biases in Investments: An Experimental Study. SSRN Working Paper 4787249.

Gao, Z., W. Jiang, and Y. Yan (2025). Detecting Lookahead Bias in LLM Forecasts. arXiv:2512.23847.

Glasserman, P. and C. Lin (2023). Assessing Look-Ahead Bias in Stock Return Predictions Generated by GPT sentiment analysis. SSRN Working Paper 4586726.

Goldsmith-Pinkham, P. and K. Shue (2023). The Gender Gap in Housing Returns. *Journal of Finance 78*(2), 1097–1145.

Grattafiori, A. et al. (2024). The Llama 3 herd of models. arXiv:2407.21783.

Grieser, W., J. A. Cookson, M. Fathollahi, and B. Venugopal (2026). LLMs and Systematic Measurement Error. SSRN Working Paper 7177500.

Hansen, A. L., J. J. Horton, S. Kazinnik, D. Puzzello, and A. Zarifhonarvar (2026). Simulating the Survey of Professional Forecasters. SSRN Working Paper 5066286.

Hansen, S., P. J. Lambert, N. Bloom, S. J. Davis, R. Sadun, and B. Taska (2023). Remote Work across Jobs, Companies, and Space. NBER Working Paper 31007.

He, S., L. Lv, A. Manela, and J. Wu (2025). Instruction Tuning Chronologically Consistent Language Models. SSRN Working Paper 5348747.

Horton, J. J., A. Filippas, and B. S. Manning (2023). Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? NBER Working Paper 31122.

Ke, S. (2024). Dissecting Misreactions in Earnings Expectations: Evidence from Analyst Reports. SSRN Working Paper 5025830.

Kojima, T., S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa (2022). Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, Volume 35, pp. 22199–22213.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology 5*(3), 213–236.

Levy, B. (2026). Caution Ahead: Numerical Reasoning and Look-Ahead Bias in AI Models. *Journal of Accounting Research 64*(3), 1139–1188.

Lopez-Lira, A., Y. Tang, and M. Zhu (2025). The Memorization Problem: Can we trust LLMs' economic forecasts? SSRN Working Paper 5217505.

Manski, C. F. (2004). Measuring expectations. *Econometrica 72*(5), 1329–1376.

Matsumoto, D. A. (2002). Management's incentives to avoid negative earnings surprises. *The Accounting Review 77*(3), 483–514.

Mincer, J. A. and V. Zarnowitz (1969). The evaluation of economic forecasts. In J. A.

Mincer (Ed.), *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*, pp. 3–46. NBER.

Ng, H., L. Peng, Y. Tao, and D. Zhou (2026). Price Discovery and Trading in Modern Prediction Markets. SSRN Working Paper 5331995.

OpenAI (2023). GPT-4 technical report. arXiv:2303.08774.

Ouyang, L., J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems (NeurIPS) 35.

Richardson, S., S. H. Teoh, and P. D. Wysocki (2004). The walk-down to beatable analyst forecasts: The role of equity issuance and insider trading incentives. *Contemporary Accounting Research 21*(4), 885–924.

Sarkar, S. K. (2024). StoriesLM: A Family of Language Models with Time-Indexed Training Data. SSRN Working Paper 4881024.

Sarkar, S. K. and K. Vafa (2024). Lookahead Bias in Pretrained Language Models. SSRN Working Paper 4754678.

Schuman, H. and S. Presser (1981). *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. New York: Academic Press.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist 54*(2), 93–105.

Tversky, A. and D. Kahneman (1981). The framing of decisions and the psychology of choice. *Science 211*(4481), 453–458.

van Binsbergen, J. H., X. Han, and A. Lopez-Lira (2023). Man versus machine learning: The term structure of earnings expectations and conditional biases. *Review of Financial Studies 36*(6), 2361–2396.

Veenman, D. and P. Verwijmeren (2018). Do investors fully unravel persistent pessimism in analysts' earnings forecasts? *The Accounting Review 93*(3), 349–377.

Wei, J., X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, Volume 35, pp. 24824–24837.

Zheng, C., H. Zhou, F. Meng, J. Zhou, and M. Huang (2024). Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*, pp. 19426–19454.

Appendix A Prompts and LLM Configuration

This appendix reproduces the verbatim prompts used in each experiment. We hold model, sampling, and output-extraction procedures fixed and vary only the prompt; the extraction procedure is described next. Bracketed placeholders `[...]` are substituted with row-specific values (firm name, ticker, evaluation date, retrieved figures). All prompts are issued to the models of Table IA1, of which LLaMA 3.3-70B is the primary model.

A.1 Models and Training Cutoffs

We use six open-weight models: the primary LLaMA 3.3-70B and, for the cross-model checks (Appendix D.3), LLaMA 3.1-8B and Qwen 2.5 at 3B, 7B, 32B, and 72B. Table IA1 records each model's inference precision and exposure-classification windows. The identification design of Appendix B is identical across models, so every cross-model test reuses these windows rather than restating them.

Table IA1: Models, Inference Precision, and Exposure-Classification Windows

Model	Params	Precision	Exposed (pre)	Unexposed (post)
LLaMA 3.3-70B†	70B	4-bit NF4	< Dec 2023	> Dec 2023
LLaMA 3.1-8B	8B	16-bit	< Dec 2023	> Dec 2023
Qwen 2.5-3B	3B	16-bit	< Jan 2024	> Aug 2024
Qwen 2.5-7B	7B	16-bit	< Jan 2024	> Aug 2024
Qwen 2.5-32B	32B	4-bit NF4	< Jan 2024	> Aug 2024
Qwen 2.5-72B	72B	4-bit NF4	< Jan 2024	> Aug 2024

Notes: † primary model. "Exposed" events resolve before the pre-cutoff date and "unexposed" after the post-cutoff date; intervening resolutions are excluded from the exposure estimation. December 2023 is the reported cutoff window for the LLaMA models. Qwen 2.5 discloses no cutoff; for those models, January–August 2024 is a conservative gray zone bounded by the September 2024 release date, not a reported training cutoff. Larger models are loaded in 4-bit NF4 quantization, smaller models in 16-bit.

The larger models (LLaMA 3.3-70B, Qwen 2.5-32B, and Qwen 2.5-72B) are loaded in 4-bit NF4 quantization, and the smaller models (LLaMA 3.1-8B, Qwen 2.5-3B, and Qwen 2.5-7B) run at 16-bit precision; loader configurations are in the replication code. Precision is therefore

confounded with model size in the cross-model comparisons. The Qwen coefficients change smoothly between 7B and 32B, but this descriptive pattern does not separately identify the effect of quantization.

A.2 Inference and Probability Extraction

Each prompt produces a single deterministic forward pass through the model in `eval()` mode under `torch.no_grad()`; we do not call `generate()` or use any sampling, so temperature, top-p, and top-k are not applied. The elicited probability comes from the model's logits at the final input position (the next-token distribution given the prompt), not from any generated text. This single-pass procedure applies to every variant in this appendix, with one exception. The chain-of-thought condition is the only departure: we first call `model.generate()` with `do_sample=False` (greedy decoding) to produce the reasoning text deterministically, then run a second forward pass on the prompt concatenated with the generated reasoning to extract the probability via the same logit-based procedure.

For each binary task we predefine the set of candidate tokens for each label, allowing for casing and leading-whitespace variants. For the prediction-market *Yes/No* extraction we use {"Yes", "yes", "Yes", "yes"} and {"No", "no", "No", "no"}; for the analyst *Increase/Decrease* extraction we use {"increase", "increase", "Increase", "Increase"} and {"decrease", "decrease", "Decrease", "Decrease"}. Logits across each token-variant set are aggregated by `logsumexp`, and a softmax over the two aggregated values returns the elicited probability $\tilde{p} \in [0, 1]$.

$\tilde{p}$ is a *conditional* probability $P(\text{label} \,|\, \text{answer} \in \{\text{Yes, No}\})$: the LLM's relative weight on the two outcome labels given that it produces one of them. We use this conditional form

rather than the marginal next-token probability $P(\text{Yes}) + P(\text{No})$ over the full vocabulary, which would instead measure how strongly the model commits to giving any directional Yes-or-No answer rather than hedging, refusing, or producing other continuations. The conditional probability is the natural counterpart to our human benchmarks (a market price or analyst-consensus probability, both elicited conditional on the human giving a directional response), produces a valid probability for every observation, and matches the inner-probability extraction in Chen et al. (2026). The sources of divergence documented in the main text are properties of this conditional probability.

Tokenizers use left-side padding (`padding_side=''left''`) so the last-position logit aligns across batched prompts, with `pad_token` set to `eos_token` when not natively defined.

A.3 Baseline Elicitations and the Persona Crowd

The prompts below are the minimal prediction-market elicitation and the informed analyst (EPS) baseline; the prediction-market baseline for the main gap results is a demographic *persona crowd* averaged over the minimal elicitation, constructed at the end of this subsection.

Prediction markets (Yes/No)

```
You are evaluating the following question using only information available up

to [date].

Question:  [contract question]

Determine whether the more likely answer is Yes or No.

You MUST respond with ONLY one word:

Yes or No

Answer:
```

Analyst earnings, informed EPS baseline (Increase/Decrease)

```
You are evaluating the following question using only information available up

to [evaluation date].

The annual EPS for [company] ([ticker]) in fiscal year [prior year] was

$[X.XX].

Most recently reported quarterly EPS:

Quarter ending [date]:  $[X.XX]

Quarter ending [date]:  $[X.XX]

Quarter ending [date]:  $[X.XX]

Quarter ending [date]:  $[X.XX]

Question:

Will the earnings per share (EPS) of [company] ([ticker]) for fiscal year

[year] increase or decrease from fiscal year [prior year]?

You MUST respond with ONLY one word:

increase or decrease

Answer:
```

Persona crowd. For the prediction-market baseline we construct a *demographic* persona crowd, prepending a descriptor-only prefix to the minimal elicitation with no roleplay header.

Following Bhagwat et al. (2026), who draw their traits from FINRA's 2021 National Financial Capability Study (NFCS), the descriptor has the fixed form

Demographic descriptor (prepended to the minimal elicitation)

```
You are a {race} {gender}, {age}, with {education}, and {income}.  You
consider yourself politically {politics}, and you have {risk}.
```

filled from a factorial grid over seven axes: five NFCS demographic items—*age* (18–34 / 35–54 / 55 or older), *gender* (male / female), *race* (White / Black / Hispanic / Asian, non-Hispanic), *household income* (< \$50K / \$50–100K / > \$100K), and *education* (some college or less / college or more)—crossed with *political orientation* (conservative / independent / liberal) and financial *risk tolerance* (low / moderate / high). The five demographic axes give 144 cells, and crossing with politics and risk yields 1,296 personas.

Eliciting all 1,296 per market-date is costly, so we stratify into 24 balanced blocks of 54: each pairs six demographic cells—a balanced partition of the 144, refined by deterministic greedy swaps—with all nine politics×risk combinations. Each market-date draws one block by the calendar rotation block = (market hash+calendar day) mod 24, so any 24 consecutive days cover the full grid and a 24-day rolling crowd mean recovers its composition. The crowd index is the equal-weighted mean of that date's 54 persona probabilities.

Shared structure. Every prompt in the subsections that follow modifies one of the baseline prompts above by a single operation: *inserting* a context line or block after the date preamble (information-provision and anchor conditions), *changing* the response labels

or their order (framing conditions), or *adding* a reasoning instruction (reasoning conditions). In prediction markets these deviation prompts modify the single minimal elicitation rather than the persona crowd. Aggregating personas does not close the gap or reverse its sign (Section 3.1), so little is lost by using the single elicitation. The effects reported below are nevertheless specific to it and need not carry over to the crowd. This choice isolates the continuation probability and avoids re-running every condition over the full persona block. In analyst earnings the deviations modify the informed EPS baseline. Everything else—the forecasting question, response instruction, answer terminator, and token-level probability extraction—is unchanged, so the subsections below state only the operation each applies.

A.4 Label-Ordering Variants

These label-ordering and wording conditions are analyzed in Section 5.3. The order-reversed variant replaces "Yes or No" with "No or Yes" (prediction markets) and "Increase or Decrease" with "Decrease or Increase" (analyst), leaving the rest of the baseline prompt unchanged. The cross-wording variant issues the analyst question under plain Yes/No labels, adjusting the phrasing to "be higher than" so that Yes/No read naturally, to test whether the directional prior is a property of the Inc/Dec wording rather than the underlying event:

Analyst earnings (Yes/No cross-wording)

```
You are evaluating the following question using only information available up
to [evaluation date].
Question:
Will the earnings per share (EPS) of [company] ([ticker]) for fiscal year
[year] be higher than in fiscal year [prior year]?
Determine whether the more likely answer is Yes or No.
You MUST respond with ONLY one word:
Yes or No
Answer:
```

The neutral alphabetic variants map the two outcomes to letters A and B with weaker outcome-specific semantic content, isolating the order effect from the labels' meaning. The favorable outcome is always labeled A; the two variants differ only in display order, listing either A or B first. The two settings share one scaffold, differing only in the preamble and the substituted outcome descriptions:

Neutral A/B, favorable first (prediction markets and analyst earnings)

Prediction-market preamble:

```
You are evaluating a binary prediction-market contract.

Contract question: [contract question]

Evaluation date: [date]. Use only information that would have been available

as of this date.
```

Analyst-earnings preamble:

```
You are evaluating the following question using only information available up

to [evaluation date].

Question: Will the earnings per share (EPS) of [company] ([ticker]) for

fiscal year [year] increase or decrease from fiscal year [prior year]?
```

Shared continuation:

```
The [contract / question] has two possible outcomes:

Outcome A: [The event described in the contract occurs / EPS increases].

Outcome B: [The event described in the contract does not occur / EPS

decreases].

Which outcome is more likely?

Answer with exactly one token: A or B.
```

A.5 Reasoning Conditions

The three reasoning conditions—baseline, reasoning instruction, and chain-of-thought—are analyzed in Section 5.4. The prediction-market prompts are shown first; the analyst variants follow, retaining the EPS context block of the informed baseline.

Reasoning Instruction (prediction markets)

```
You are evaluating the following question using only information available up

to [date].

Question:  [contract question]

Think through this question step by step.

Determine whether the more likely answer is Yes or No.

You MUST respond with ONLY one word:

Yes or No

Answer:
```

Chain-of-Thought (two-pass). Pass 1 elicits reasoning without a verdict:

Chain-of-Thought, Pass 1

```
You are evaluating a question using only information available up to [date].

Take me through the relevant facts, uncertainties, and possibilities step by

step.

Your goal is to determine which answer (Yes or No) would be more likely, but

do NOT state a final Yes or No.

Do NOT include a final verdict or any explicit 'Yes'/'No' answer.

Question:  [contract question]

Notes:
```

Pass 2 appends the generated reasoning trace and extracts the verdict:

Chain-of-Thought, Pass 2

```
You are evaluating the following question using only information available up
to [date].
Question: [contract question]
Determine whether the more likely answer is Yes or No.
If the outcome is uncertain, choose the option that seems more likely to be
true.
Notes: [generated reasoning]
Respond with ONLY one word:
Yes or No
Answer:
```

In the analyst variants below, the `[EPS context block]` placeholder is the prior annual and quarterly EPS lines of the informed baseline above.

Reasoning Instruction (analyst earnings)

```
You are evaluating the following question using only information available up
to [evaluation date].
```

[EPS context block]

```
Question:
Will the earnings per share (EPS) of [company] ([ticker]) for fiscal year
[year] increase or decrease from fiscal year [prior year]?
Think through this question step by step.
Determine whether the more likely answer is increase or decrease.
You MUST respond with ONLY one word:
increase or decrease
Answer:
```

Chain-of-Thought, Pass 1 (analyst earnings)

```
You are evaluating a question using only information available up to
[evaluation date].
```

[EPS context block]

```
Take me through the relevant facts, uncertainties, and possibilities step by
step.
Your goal is to determine which answer (increase or decrease) would be more
likely, but do NOT state a final increase or decrease.
Do NOT include a final verdict or any explicit ‘increase’/‘decrease’ answer.
Question:  [analyst baseline question]
```

Chain-of-Thought, Pass 2 (analyst earnings)

```
You are evaluating the following question using only information available up
to [evaluation date].
```

[EPS context block]

```
Question: [analyst baseline question]
The following is a step-by-step analysis of this question:
Reasoning: [generated reasoning]
Based on the reasoning above, determine whether the more likely answer is
increase or decrease.
If the outcome is uncertain, choose the option that seems more likely to be
true.
Respond with ONLY one word:
increase or decrease
Answer:
```

A.6 Context Provision (RAG)

In prediction markets, the previous-day condition (Section 5.2) inserts one line after the `Question:` line of the baseline prompt above, before the "Determine whether" instruction: "As of [previous date], the prediction market price for this contract is [price], reflecting market participants' current probability estimate."

In analyst earnings, each disclosure condition inserts a block after the date preamble, before the `Question:` line, leaving the rest of the informed EPS baseline prompt unchanged: "The following is an excerpt from a financial filing for [company] ([ticker]): [excerpt]". The `[excerpt]` is the condition's selected text (the four conditions and their sentence selection

are defined in Section 2), front-truncated to a 5,000-token budget when needed and drawn from the most recently available filing.

A.7 Prediction-Market Category Construction

A Kalshi market inherits the category of its *series*: the exchange assigns each series ticker one of fourteen native values, which we map to ten labels (Politics, Sports, Entertainment, Science & Technology, Climate & Weather, Crypto, Economics, Financials, Commodities, and Companies). Four of the fourteen—Economics, Financials, Commodities and Companies—are already among the ten and are kept as they are; the other ten native values are grouped into the remaining six labels.

Polymarket markets carry tags of their own, but they are assigned by hand and much noisier than Kalshi's series categories: a market carries a median of four, and only 5,383 of the 7,208 resolve to a single one of the ten labels. We therefore assign each Polymarket question by nearest centroid on unit-normalized question embeddings (OpenAI `text-embedding-3-small`), in two stages. The first takes the nearest of seven top-level centroids (Politics, Finance & Economics, Crypto, Science & Technology, Climate & Weather, Sports, Entertainment); where that is Finance & Economics, the second resolves it against four finance sub-centroids (Economics, Financials, Commodities, Companies), which recovers the ten labels. The centroids are trained on the exchange-labelled questions.

The tags serve only as a check. On the 5,383 markets whose tags do resolve, the embedding assignment agrees with the mapped tag 89% of the time. Remaining classification error can mix categories and attenuate heterogeneity, although the category labels do not enter prompting or exposure assignment.

For analysis the ten labels are reduced to eight. Commodities and Companies are too sparse to estimate on their own (34 and 10 markets), so they are pooled with Financials into a single *Financials, Commodities & Companies* category. Politics is the omitted reference.

Appendix B Cutoff-Based Identification and Validity Checks

For LLaMA 3.3-70B (the primary model), the training-corpus exposure effect $\hat{\lambda}$ of Section 4.1 classifies observations resolving before December 1, 2023 as pre-cutoff and those resolving after December 31, 2023 as post-cutoff, excluding December 2023, the reported training-cutoff month. This appendix defends that identification: we lay out the formal potential-outcomes setup and the assumptions under which $\hat{\lambda}$ recovers the average effect of pre-cutoff exposure, then stress the design directly: balance evidence on the task-difficulty confound, a donut specification, a placebo cutoff, and the effect's sensitivity to temporal distance from the cutoff.

B.1 Framework: Assumptions on the Training Cutoff

Let $p_{it}^{\mathrm{LLM}}(d)$ denote the potential outcome-aligned LLM probability for task i at time t under exposure status $d \in \{0, 1\}$. Because the human benchmark for a fixed task-date does not change with the model's training exposure, define the potential signed gap as $s_{it}(d) = p_{it}^{\mathrm{LLM}}(d) - p_{it}^{H}$. The task-level exposure effect is therefore

$$s_{it}(1) - s_{it}(0) = p_{it}^{\mathrm{LLM}}(1) - p_{it}^{\mathrm{LLM}}(0).$$

We interpret $\lambda \equiv \mathbb{E}[s_{it}(1) - s_{it}(0)]$ as the average effect of eligibility for training-corpus

exposure under the following conditions:

Assumption B1 (Credibly bounded training horizon)**.** *The LLM's training corpus contains no material generated after T_1, the close of the reported cutoff window. Cutoff-side assignment d_i is determined by resolution timing and creates a valid contrast in eligibility for outcome-revealing corpus exposure.*

Assumption B2 (Conditional exchangeability and overlap)**.** *Conditional on observed controls Γ_{it}:*

$$(s_{it}(1),\, s_{it}(0)) \perp d_i \mid \Gamma_{it}, \qquad 0 < \Pr(d_i = 1 \mid \Gamma_{it}) < 1.$$

Consistency additionally requires the observed gap to equal $s_{it}(d_i)$. Under consistency, conditional exchangeability, overlap, and a correctly specified conditional-mean model, the adjusted cross-cutoff contrast identifies λ. Because some eligible pre-cutoff outcomes may never enter the corpus, λ is best interpreted as the effect of eligibility for exposure rather than the effect of known inclusion in the training data.

The primary threat to the conditional exchangeability assumption is that tasks resolving before versus after the cutoff differ in inherent difficulty independently of training-corpus exposure. Two pieces of evidence address this. First, the controls Γ_{it} adjust for observed differences in task characteristics (market or sector composition, analyst coverage, and firm fundamentals where applicable). Second, the balance evidence below bears directly on this threat.

B.2 Balance Evidence: Outcome-Aligned Human Probabilities

Table IA2 shows that exposed and unexposed markets differ on observables—pre-cutoff markets are shorter-duration, substantially thinner-traded, and more Kalshi-weighted—so the two groups are not identically distributed. The average outcome-aligned human probability is also lower before the cutoff (0.760 vs. 0.830), the opposite direction from an easier-pre-cutoff-task explanation. Table IA3 shows that this difference survives adjustment for observed imbalances: the exposure coefficient ranges from −0.029 to −0.065. These results argue against that particular difficulty confound, although they do not establish conditional exchangeability with respect to all unobserved determinants of the gap.

Table IA2: Pre- and Post-Cutoff Prediction-Market Characteristics

Variable	Pre-cutoff ($N = 815$)	Post-cutoff ($N = 7,316$)	Difference (SE)	p-value
Outcome, P(YES)	0.352	0.222	0.130^{***} (0.017)	<0.001
Duration (days)	44.4	57.3	-13.0^{***} (2.2)	<0.001
Log volume (within-source z-score)	−0.619	0.069	-0.688^{***} (0.039)	<0.001
Mean human price, P(correct outcome)	0.760	0.830	-0.070^{***} (0.008)	<0.001
Source: Polymarket (=1)	0.557	0.923	-0.366^{***} (0.018)	<0.001

Notes: Unit of observation is the prediction-market question, on the canonical prediction-market panel (single-leg contracts, at least seven days in duration), 8,131 markets. Pre-cutoff markets resolve before December 2023 (the LLaMA 3.3 70B cutoff), post-cutoff after; gray-zone markets are excluded. Log volume is standardized within platform. The price row is the market's mean outcome-aligned price across the dates it trades inside the 30-day window, the same panel-level human benchmark the placebo regression of Table IA3 takes as its dependent variable, rather than a single-date snapshot. Standard errors of the difference are from two-sample Welch t-tests. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

B.3 Donut Specification

Observations resolving close to the training cutoff are the hardest to classify cleanly: their outcomes may have entered the training corpus partially or with low coverage density, so a

Table IA3: Difficulty Placebo: Outcome-Aligned Human Probability

	(1) Baseline	(2) + Source	(3) + Vol/Dur	(4) + Cal FE
Dependent variable	Human market price, P(correct outcome)			
Exposure	−0.065***	−0.052***	−0.031***	−0.029***
	(0.009)	(0.010)	(0.009)	(0.009)
Source FE	No	Yes	Yes	Yes
Log volume, log duration	No	No	Yes	Yes
Calendar day FE	No	No	No	Yes
Observations	158,814	158,814	158,410	158,410
Clusters (markets)	8,131	8,131	8,131	8,131
R^2	0.007	0.010	0.063	0.075

Notes: The dependent variable is the human market price transformed to the probability assigned to the realized outcome, on the canonical prediction-market panel. Exposure equals one for markets resolving before December 2023; a negative coefficient means that the human benchmark assigns less probability to the realized outcome before the cutoff. Standard errors are two-way clustered by market and evaluation date, as in the other prediction-market tables. Both clustering dimensions are large here (8,131 markets and 1,382 evaluation dates), so the Cameron–Gelbach–Miller covariance is positive semidefinite and no eigenvalue adjustment is applied. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

natural concern is that misclassification in this ambiguous window inflates $\hat{\lambda}$ by mixing genuinely exposed and unexposed observations near the boundary. To eliminate any plausibly ambiguous observation, we re-estimate the signed-gap directional regression of Equation 3 on the prediction-market sample after dropping all markets resolving within a donut window of w months on either side of the December 2023 cutoff, for $w \in \{3, 6, 9, 12\}$; standard errors are two-way clustered by market and evaluation date. Table IA4 reports the results. The exposure × outcome interaction $\hat{\theta}$ strengthens monotonically as the donut widens (0.191 with no exclusion, rising to 0.424 at ±12 months in the fully controlled column), arguing against near-cutoff misclassification as the source of the directional asymmetry documented in Section 4.2. The pooled exposure coefficient $\hat{\lambda}$ is positive in every window and significant at the 5% level in all but the widest (0.040 with no exclusion, rising to a peak of 0.076 at ±9 months and 0.072 (SE 0.037) at ±12, where it is significant only at the 10% level and the treated sample has fallen from 815 to 275 markets). Excluding the ambiguous boundary therefore strengthens rather than eliminates $\hat{\lambda}$, and leaves $\hat{\theta}$ intact.

Table IA4: Training-Corpus Exposure across Donut Exclusion Windows (Signed Gap)

	(1)	(2)	(3)	Markets	
Exclusion window	Baseline	+Source	+Full Controls	N_{treated}	N_{total}
Main specification (no exclusion)					
Exposure	0.050***	0.061***	0.040**	815	8,131
	(0.016)	(0.017)	(0.017)		
Exposure × Outcome	0.190***	0.189***	0.191***		
	(0.033)	(0.033)	(0.033)		
Donut ±3 *months*					
Exposure	0.071***	0.089***	0.067***	656	7,620
	(0.018)	(0.020)	(0.020)		
Exposure × Outcome	0.215***	0.213***	0.214***		
	(0.036)	(0.036)	(0.036)		
Donut ±6 *months*					
Exposure	0.073***	0.100***	0.073***	499	6,837
	(0.021)	(0.024)	(0.024)		
Exposure × Outcome	0.282***	0.276***	0.281***		
	(0.039)	(0.039)	(0.039)		
Donut ±9 *months*					
Exposure	0.055**	0.092***	0.076**	350	5,587
	(0.024)	(0.031)	(0.031)		
Exposure × Outcome	0.347***	0.342***	0.339***		
	(0.043)	(0.043)	(0.042)		
Donut ±12 *months*					
Exposure	0.029	0.084**	0.072*	275	4,002
	(0.027)	(0.037)	(0.037)		
Exposure × Outcome	0.438***	0.427***	0.424***		
	(0.047)	(0.047)	(0.046)		

Notes: The dependent variable is the signed LLM–human gap s_{it} on the canonical prediction-market panel, under the demographic persona-crowd baseline (matching Table 4 Panel A). The Exposure row reports the pooled exposure coefficient; the Exposure × Outcome row reports the directional asymmetry $\hat{\theta}$ from the specification adding the outcome and exposure × outcome terms. Column (3) uses the full column-(4) control set: source FE, log volume, and calendar-week fixed effects. Each row excludes markets resolving within $\pm w$ months of the December 2023 training cutoff. Standard errors are two-way clustered by market and evaluation date. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

B.4 Placebo Cutoff

Two threats to identification motivate this test. First, the exposure effect we estimate could reflect a general time trend rather than the training cutoff itself: early-resolving markets might be systematically easier or have different composition. Second, the effect could be smooth across time rather than sharp at December 2023, in which case the cutoff date carries no special status. To address both threats, we re-estimate the signed-gap directional

regression of Equation 3 on the prediction-market sample, reassigning the treatment indicator d_i to placebo cutoff dates spanning June 2022 through December 2024 in place of the true December 2023 cutoff; standard errors are two-way clustered by market and evaluation date. Table IA5 reports the placebo coefficients alongside the true-cutoff estimate.

The pre-true-cutoff placebos (June and December 2022) test the first threat: a generic time trend would survive substitution at any date. These are not clean placebos, because reassigning the cutoff to 2022 leaves genuinely exposed markets (those resolving up to December 2023) in the control group while classifying the oldest, most-exposed markets as treated. With full controls in column (3) they accordingly retain a positive coefficient (+0.050 and +0.066), comparable to the true-cutoff effect of +0.040, since their treated group is itself dominated by genuinely exposed markets—so these are weaker variants of the true split, not valid null tests.

The post-true-cutoff placebos (June and December 2024) provide the cleaner test, since they relabel only genuinely unexposed post-cutoff months into the treated group. The June 2024 placebo falls to −0.007 and is statistically indistinguishable from zero, and the December 2024 placebo is significantly negative (−0.050). These estimates are consistent with a localized contrast near the reported cutoff rather than a smooth trend, although each placebo changes the treated–control composition. The sample imbalance (815 pre-cutoff vs. 7,316 post-cutoff markets, a 1:9 ratio) means each placebo generates a very different treated/control split, so the coefficients should be read as a battery of complementary tests.

Table IA5: Training-Corpus Exposure at Placebo Cutoffs (Signed Gap)

	(1)	(2)	(3)	Markets	
Assigned cutoff	Baseline	+Source	+Full Controls	N_{treated}	N_{total}
True cutoff (Dec 2023)	0.050***	0.061***	0.040**	815	8,131
	(0.016)	(0.017)	(0.017)		
June 2022	0.027	0.058*	0.050	242	8,119
	(0.028)	(0.034)	(0.033)		
December 2022	0.036	0.072**	0.066**	274	8,128
	(0.026)	(0.032)	(0.032)		
June 2024	0.015	0.017	−0.007	1,709	7,817
	(0.013)	(0.013)	(0.014)		
December 2024	−0.048***	−0.047***	−0.050***	4,263	7,438
	(0.010)	(0.011)	(0.010)		
Source FE	No	Yes	Yes		
Volume	No	No	Yes		
Calendar FE	No	No	Yes		

Notes: The dependent variable is the signed LLM–human gap s_{it} on the canonical prediction-market panel, under the demographic persona-crowd baseline (matching Table 4 Panel A). Each row reassigns the cutoff to a counterfactual date, with Exposure equal to one for markets resolving before that date. Column (3) uses the full column-(4) control set: source FE, log volume, and calendar-week fixed effects. Standard errors are two-way clustered by market and evaluation date. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

B.5 Distance to the Training Cutoff

Exposure could shift the gap in two ways: a discrete jump at the cutoff, and a gradient in how far an event resolved from it. A distance gradient is a natural concern on both sides. Before the cutoff, an event that resolved longer ago has had more time for post-resolution coverage to accumulate and be absorbed into training, so one might expect it to be more exposed and its gap more narrowed the further back it resolved; after the cutoff, the frozen model may instead fall further behind an accumulating human as the forecast date recedes, widening the gap. Because either gradient could equally reflect a calendar trend unrelated to training, separating it from the discrete jump is what isolates the memorization signal. We instantiate Equation 2 with $Z_{it} = \text{distance}_i$, the number of months between event resolution and the nearest edge of the December 2023 cutoff window, so it is zero where exposure

switches and always non-negative:

$$s_{it} = \alpha + \lambda d_i + \chi \, \text{distance}_i + \pi \, (d_i \times \text{distance}_i) + \Gamma_{it}' \eta + \varepsilon_{it}.$$

The analyst columns carry analyst coverage, event-time effects, and the four firm controls of Section 2; the prediction-market columns carry source fixed effects and event-time effects. Adding the firm controls leaves the distance terms unchanged.

Table IA6: Signed LLM–Human Gap and Distance to Training Cutoff

	Panel A: Prediction Markets		Panel B: Analyst Earnings	
	(1) Baseline	(2) Full	(3) Baseline	(4) Full
Exposure (λ)	0.0933***	0.0877***	0.0815***	0.0867***
	(0.027)	(0.027)	(0.024)	(0.029)
Distance to Cutoff (χ, months)	0.0038***	0.0039***	0.0025***	0.0024**
	(0.001)	(0.001)	(0.000)	(0.001)
Distance × Exposure (π)	−0.0041*	−0.0021	−0.0017**	−0.0022
	(0.002)	(0.002)	(0.001)	(0.001)
Log Active Analysts				0.0025
				(0.012)
Source Controls	No	Yes	No	No
Firm Controls	No	No	No	Yes
Calendar FE	No	Yes	No	Yes
Observations	158,814	158,814	132,957	132,580
R^2	0.003	0.006	0.006	0.015

Notes: Dependent variable: signed gap $s_{it} = \alpha + \lambda d_i + \chi \, \text{dist}_i + \pi \, (d_i \times \text{dist}_i) + \Gamma_{it}' \eta + \varepsilon_{it}$. Distance is months from resolution to the December 2023 cutoff; χ is the common post-cutoff slope, π the additional pre-cutoff slope, the pre-cutoff gradient $\chi + \pi$, and λ the level shift at the cutoff. Calendar FE are event-time fixed effects. Analyst panel requires ≥ 5 active analysts. SEs two-way clustered by market and evaluation date (PM), firm and fiscal year (analyst). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

The level shift at the cutoff, $\hat{\lambda}$, is large and significant in both settings (+8.8 to +9.3 pp in prediction markets, +8.2 to +8.7 pp in analyst earnings), but the distance terms carry no exposure signal: the exposed-side interaction $\hat{\pi}$ is small and, with controls, insignificant, and it roughly offsets the common (post-cutoff) per-month slope $\hat{\chi}$, so the pre-cutoff per-month slope $\hat{\chi} + \hat{\pi}$ is statistically indistinguishable from zero in both settings. Relative to the common calendar trend, the gap does not narrow further with how long before the cutoff

an outcome resolved; the memorization signal is the discrete shift $\hat{\lambda}$ at the cutoff, not a distance gradient. The positive post-cutoff slope $\hat{\chi}$ is also the opposite of what a post-cutoff information deficit would imply, under which the frozen model would fall further behind the accumulating human as distance grows. The main text rests on the estimated discontinuous shift $\hat{\lambda}$, which a smooth calendar trend cannot mimic; a discontinuous omitted factor at the same date remains a possible confound, and the per-month gradient cannot be cleanly separated from a calendar trend in this specification.

Appendix C Robustness of the Main Results

This appendix re-estimates the paper's headline results in alternative samples and specifications. It considers a different macroeconomic regime in the analyst panel and stricter truncation of the disclosure excerpts; the principal signs and structure of the LLM–human gap remain similar.

C.1 Robustness to the Macro Regime (Analyst Panel)

For LLaMA 3.3-70B the analyst exposure split places fiscal years 2020–2022 before the training cutoff and 2024–2025 after it, so the pre-cutoff window coincides with the COVID-19 period and its recovery. A natural concern is that the exposure effect reflects a difference in earnings predictability across macro regimes rather than training-corpus recall. We address this by decomposing the gap by fiscal year and re-estimating the exposure narrowing after dropping the COVID years.

Table IA7 reports the mean signed gap by fiscal year. Under the informed EPS baseline

the exposed years carry similar gaps, with no single year driving the narrowing; in particular the fiscal-2020 COVID year is not an outlier. Restricting the exposed group to fiscal 2022, a post-COVID but still pre-cutoff year, and comparing it against the post-cutoff years (2024–2025) gives a post-acute-COVID comparison.

Table IA7: Mean Signed Gap by Fiscal Year, Analyst Earnings

Fiscal year	Regime	Mean signed gap (pp)
2020	Exposed (COVID)	−14.6
2021	Exposed	−11.2
2022	Exposed	−16.6
2024	Unexposed	−21.3
2025	Unexposed	−18.4

Notes: Unconditional mean signed gap $s_{it} = p_{it}^{\text{LLM}} - p_{it}^{H}$ for the analyst panel, by fiscal year (firm-year-date observations, no controls). Fiscal year 2023 is excluded as the cutoff year. Exposed years (2020–2022) resolve before the December 2023 cutoff; unexposed years (2024–2025) resolve after.

Restricting the exposed group to fiscal 2022 preserves the sign of the exposure narrowing but attenuates its magnitude. The result shows that the COVID years do not fully account for the cross-cutoff difference, although it does not eliminate all macro-regime confounding. These are unconditional mean differences; the controlled $\hat{\lambda}$ refines the magnitudes but not the comparison.

C.2 Runtime Truncation Robustness

The 10-K disclosure conditions are front-truncated to a 5,000-token prompt budget when needed (Appendix A, "Context Provision (RAG)"); the truncation rate in the stacked analysis sample is 46.3% for Sampled MD&A, 45.3% for Truncated MD&A, 27.8% for Risk Factors, and 17.0% for Forward-Looking Statements. This subsection tests whether head-truncation independently moves the signed analyst gap, which would confound the information-provision results of Section 5.1. We add a head-truncation indicator—equal to one when the LLM run for a (firm-year, disclosure) cell front-truncated the disclosure text—to the regression speci-

fication of Table 6 (column 3, on the Yearly-plus-Quarterly EPS baseline), and build up the controls across columns.

Table IA8 reports the head-truncation coefficient across the build-up. It is small and statistically indistinguishable from zero throughout (-0.014 to -0.015, all $p > 0.5$) and stable as firm controls and disclosure $\times$ exposure interactions are added. We therefore find no detectable average truncation effect in these specifications. The cautionary sections associated with the disclosure flip (Risk Factors and Forward-Looking Statements) are also the least frequently truncated (28% and 17%, compared with 45–46% for the neutral MD&A excerpts), which further argues against truncation as the sole explanation.

Table IA8: Signed LLM–Analyst Gap: Head-Truncation Robustness

	(1) Core	(2) + Firm controls	(3) + Context $\times$ Exposure
Head-truncated disclosure	-0.0148	-0.0149	-0.0144
	(0.0227)	(0.0296)	(0.0300)
Regression core (Table 6 col. 3)	Yes	Yes	Yes
Firm controls & days-to-resolution FE	No	Yes	Yes
Context $\times$ Exposure (μ_k)	No	No	Yes
Observations	673,174	653,933	653,933
R^2	0.111	0.115	0.116

Notes: The head-truncation indicator equals one when the LLM run for a (firm-year, disclosure) cell front-truncated the disclosure to the 5,000-token prompt budget. Its coefficient is the signed LLM–analyst gap for head-truncated cells net of the disclosure and directional structure of Table 6 (column 3, stacked on the Yearly+Quarterly EPS baseline). Column (1) adds the indicator to that specification; column (2) adds firm controls and days-to-resolution (weekly) fixed effects; column (3) adds the disclosure $\times$ exposure interactions. Sample: S&P 500 firm-years with at least five active analysts; standard errors two-way clustered by company and fiscal year. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. Estimated on the 2,789 firm-years described in the note to Figure 4.

C.3 Held-Out Validation and Bounded Corrections

Section 4.3 uses the linear subtraction because it maps the estimated gap components transparently into the downstream regressions. This table evaluates whether that subtraction also recovers the individual human probabilities on firms excluded from estimation. It compares

the raw LLM probability with the ex ante linear score, that score clipped to $[0, 1]$, and an ex ante log-odds correction that is bounded by construction. The ex post clipped score adds g_i and $d_i \times g_i$ and is an infeasible benchmark. We report both the ex post linear score used in Table 5 and its clipped counterpart.

Table IA9: Held-Out Recovery of the Human Probability

Series	Bias	MAE	RMSE	Correlation	Cal. intercept	Cal. slope
Raw LLM probability	−0.117	0.283	0.408	0.522	0.345	0.581
Ex ante linear score	+0.005	0.283	0.381	0.547	0.255	0.609
Ex ante linear score, clipped	−0.024	0.253	0.367	0.551	0.227	0.681
Ex ante log-odds probability	+0.043	0.223	0.361	0.581	0.177	0.688
Ex post linear score	+0.004	0.267	0.353	0.661	0.240	0.633
Ex post linear score, clipped	−0.039	0.212	0.329	0.666	0.195	0.749

Notes: All statistics are computed on the raw increase-probability scale over the correction sample (132,580 observations, 578 companies, 2,644 firm-years), with two-fold cross-fitting by firm. Bias is candidate minus human; MAE and RMSE are the corresponding individual forecast losses. Bias is therefore not the outcome-aligned mean gap $\bar{s}$ reported in Table 5; outcome alignment leaves MAE and RMSE unchanged but can change a signed mean. Calibration intercept and slope come from regressing the human probability on the candidate series. The ex ante models use exposure, firm fundamentals, sector effects, exposure × sector effects, and forecast-week effects. The log-odds variant fits the LLM–human gap after a logit transform and maps the corrected value back through the logistic function. The ex post model additionally uses the realized outcome and its interaction with exposure. The ex ante and ex post linear scores fall outside $[0, 1]$ for 28.2% and 40.0% of observations, respectively. Clipping the ex post score changes its outcome-aligned mean gap from −0.2 to −5.6 percentage points while reducing MAE from 0.267 to 0.212. The ex ante model's held-out R^2 for the raw-scale gap is 0.051; the corresponding ex post value is 0.183.

Appendix D Additional Results

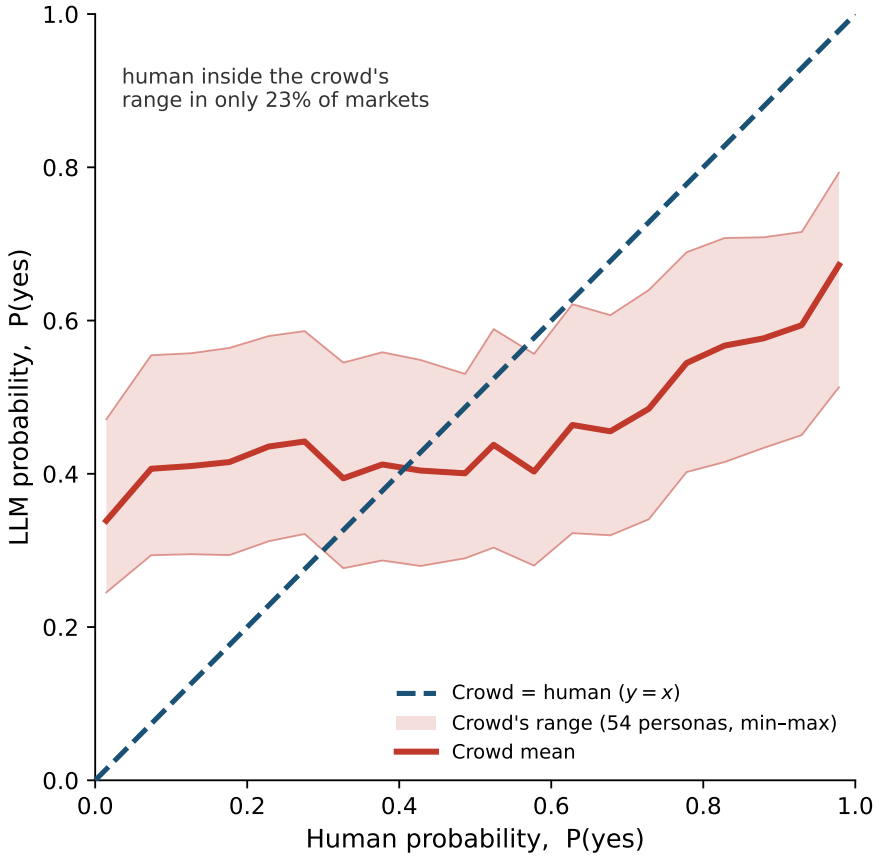
This appendix collects supplementary analyses and full estimates that extend the main argument without being central to it: the persona crowd's compression against the market price, category and sector heterogeneity, the cross-model replication, and the baseline label-ordering flip test.

D.1 Persona-Crowd Coverage of the Market Price

Figure IA1 plots how the demographic persona crowd (constructed in Appendix A.3) behaves relative to the human price, the basis for the compression discussion in Section 3.1. The human price falls within the crowd's range in only 23% of markets, and the crowd mean regresses toward 0.5 rather than following the market to its confident extremes.

Figure IA1: Persona-Crowd Forecasts Against the Market Price

Alt text: A square panel with the human probability running left to right and the model probability bottom to top, and a dashed diagonal marking equality. A shaded band shows the range across the fifty-four personas and a solid line their mean. The mean line is far flatter than the diagonal, rising only from about one third to about two thirds as the human probability sweeps the full range. A note reports that the human value falls inside the crowd range for twenty-three percent of markets.



Notes: The demographic persona crowd's forecast is plotted against the human market price, with the shaded band showing the range across personas and the solid line the crowd mean; the 45° line marks agreement with the human.

D.2 Category and Sector Heterogeneity

This subsection collects the full regression estimates underlying Figure 5 in the main text: the prediction-market topic-category specification (Table IA10) and the analyst-earnings GICS-sector specification (Table IA11). Unlike analyst earnings, where GICS classifies every firm, prediction-market category labels come from the Kalshi-based classifier described in Appendix A.7 and are applied uniformly to both platforms.

D.3 Cross-Model Extensions

This subsection replicates the headline results for the full model grid in both settings, under the identification convention of Appendix B and the model windows of Table IA1, with the paper's informed baselines, column-(4) controls, and two-way clustering throughout. The directional table below pairs prediction markets (Panel A) with analyst earnings (Panel B).

Table IA12 reports the full directional specification of Equation 3 for each model, with the realized-direction and topic-or-sector heterogeneity terms and the same column-(4) controls; the LLaMA 3.3-70B column reproduces the primary-model estimates of Table 4.

D.4 Baseline Label-Ordering Flip Test

Figure IA2 is the baseline flip test of Section 5.3: for each event, the elicited probability under one label ordering is plotted against the probability under the reversed ordering, with the training-exposure component removed. Points off the 45° line move with prompt form, and the red points—those crossing the 0.5 boundary—are the events whose directional call flips, at the per-framing rates reported in Section 5.3 and on the panel labels.

Table IA10: Topic Heterogeneity in the Pre-Cutoff Exposure Effect

	(1) Baseline	(2) +Source	(3) +Volume	(4) +Calendar FE
Pre-Cutoff Exposure (Politics baseline)	0.0615***	0.0531**	0.0487*	0.0475*
	(0.023)	(0.025)	(0.025)	(0.025)
Category (base = Politics)				
Economics	−0.1108***	−0.1145***	−0.1159***	−0.1159***
	(0.030)	(0.030)	(0.030)	(0.030)
Crypto	0.0430**	0.0448**	0.0453**	0.0450**
	(0.018)	(0.018)	(0.018)	(0.018)
Financials, Commodities & Companies	0.0006	−0.0017	−0.0029	0.0016
	(0.032)	(0.032)	(0.032)	(0.032)
Science & Technology	−0.0929***	−0.0936***	−0.0972***	−0.0974***
	(0.027)	(0.026)	(0.026)	(0.026)
Climate & Weather	0.0077	0.0077	0.0075	0.0098
	(0.042)	(0.042)	(0.042)	(0.042)
Sports	0.1865***	0.1885***	0.1830***	0.1850***
	(0.011)	(0.012)	(0.012)	(0.012)
Entertainment	0.0800***	0.0793***	0.0758***	0.0792***
	(0.018)	(0.018)	(0.019)	(0.019)
Pre-Cutoff Exposure × Category				
Economics	0.1729***	0.1697***	0.1701***	0.1723***
	(0.057)	(0.056)	(0.056)	(0.056)
Crypto	−0.0086	0.0000	−0.0022	0.0010
	(0.055)	(0.055)	(0.055)	(0.055)
Financials, Commodities & Companies	−0.0088	−0.0077	−0.0106	−0.0060
	(0.085)	(0.084)	(0.084)	(0.085)
Science & Technology	0.1689***	0.1642***	0.1675***	0.1730***
	(0.050)	(0.050)	(0.050)	(0.050)
Climate & Weather	−0.1527	−0.1588	−0.1598	−0.1621
	(0.105)	(0.105)	(0.105)	(0.106)
Sports	0.0400	0.0484	0.0462	0.0387
	(0.054)	(0.054)	(0.054)	(0.053)
Entertainment	−0.0524	−0.0456	−0.0452	−0.0474
	(0.058)	(0.058)	(0.058)	(0.057)
Source FE	No	Yes	Yes	Yes
Volume	No	No	Yes	Yes
Calendar FE	No	No	No	Yes
Observations	158,814	158,814	158,410	158,410
Clusters	8,131	8,131	8,131	8,131
R^2	0.055	0.056	0.055	0.058

Notes: Dependent variable: the outcome-aligned signed gap at the persona-crowd baseline; Exposure is one before the December 2023 cutoff and Politics is omitted, so a category's total effect is Exposure plus its interaction (stars test the difference from Politics). Categories follow Kalshi's taxonomy; Polymarket markets are assigned by embedding similarity to exchange-labeled centroids (Appendix A.7). SEs two-way clustered by market and date. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table IA11: Sector Heterogeneity in the Pre-Cutoff Exposure Effect

	(1) Baseline	(2) +Firm Ctrl	(3) +Coverage	(4) +Weeks FE
Pre-Cutoff Exposure (Industrials baseline)	0.0424	0.0421	0.0409	0.0407
	(0.032)	(0.031)	(0.029)	(0.030)
Sector (base = Industrials)				
Energy	−0.1877***	−0.2055***	−0.2032***	−0.2047***
	(0.055)	(0.053)	(0.053)	(0.053)
Materials	−0.0126	−0.0241	−0.0234	−0.0235
	(0.017)	(0.017)	(0.016)	(0.017)
Consumer Discretionary	−0.1031**	−0.1057**	−0.1023**	−0.1037**
	(0.044)	(0.046)	(0.047)	(0.047)
Consumer Staples	−0.1363***	−0.1413***	−0.1405***	−0.1406***
	(0.039)	(0.038)	(0.038)	(0.039)
Health Care	−0.0402***	−0.0458***	−0.0450***	−0.0445***
	(0.010)	(0.007)	(0.007)	(0.008)
Financials	−0.0051	−0.0270	−0.0313	−0.0286
	(0.052)	(0.053)	(0.053)	(0.052)
Information Technology	0.0028	−0.0019	0.0020	0.0004
	(0.040)	(0.039)	(0.042)	(0.041)
Communication Services	−0.0436	−0.0569	−0.0539	−0.0547
	(0.046)	(0.048)	(0.049)	(0.049)
Utilities	−0.1295**	−0.1427**	−0.1523**	−0.1461**
	(0.064)	(0.066)	(0.069)	(0.063)
Real Estate	−0.0002	−0.0052	−0.0160	−0.0109
	(0.035)	(0.036)	(0.044)	(0.035)

Notes: This panel reports the *Pre-Cutoff Exposure* baseline and the *Sector* main effects; the *Pre-Cutoff Exposure* × *Sector* interactions and the fit statistics continue in the next table, where the full notes appear.

Table IA11: Sector Heterogeneity in the Pre-Cutoff Exposure Effect, continued

	(1) Baseline	(2) +Firm Ctrl	(3) +Coverage	(4) +Weeks FE
Pre-Cutoff Exposure × *Sector*				
Energy	0.2530***	0.2454***	0.2496***	0.2473***
	(0.068)	(0.061)	(0.064)	(0.061)
Materials	−0.0278	−0.0265	−0.0252	−0.0262
	(0.037)	(0.033)	(0.031)	(0.033)
Consumer Discretionary	0.0714	0.0724	0.0728	0.0720
	(0.047)	(0.051)	(0.050)	(0.051)
Consumer Staples	0.0481	0.0487	0.0489	0.0489
	(0.084)	(0.083)	(0.082)	(0.083)
Health Care	−0.0614	−0.0602	−0.0617	−0.0607
	(0.074)	(0.070)	(0.073)	(0.071)
Financials	−0.0278	−0.0383	−0.0363	−0.0378
	(0.057)	(0.059)	(0.056)	(0.058)
Information Technology	−0.0469	−0.0478	−0.0489	−0.0483
	(0.081)	(0.078)	(0.080)	(0.080)
Communication Services	0.0047	0.0055	0.0061	0.0056
	(0.053)	(0.053)	(0.053)	(0.053)
Utilities	−0.0624	−0.0616	−0.0598	−0.0622
	(0.097)	(0.095)	(0.092)	(0.095)
Real Estate	0.0646	0.0623	0.0642*	0.0639
	(0.040)	(0.039)	(0.039)	(0.039)
Firm Controls	No	Yes	Yes	Yes
Analyst Coverage	No	No	Yes	Yes
Weeks-to-Resolution FE	No	No	No	Yes
Observations	132,957	132,580	132,580	132,580
R^2	0.030	0.033	0.034	0.038

Notes: One interacted regression per column of the signed LLM–human gap $s_{it} = p_{it}^{\text{LLM}} - p_{it}^{H}$ on pre-cutoff exposure, GICS-sector indicators, and their interaction (analyst-earnings sample, LLaMA 3.3-70B, EPS-informed baseline). Industrials is the omitted sector, so the *Pre-Cutoff Exposure* row is the exposure effect for Industrials; the *Sector* block gives each sector's gap relative to Industrials among post-cutoff firm-years; and the *Exposure* × *Sector* block gives each sector's additional exposure effect, so a sector's net exposure effect is the sum of the *Pre-Cutoff Exposure* row and its interaction. This is the analyst counterpart to Table IA10. This table is estimated on the canonical analyst sample of Section 2 and is two-way clustered, matching every other analyst exhibit; earlier versions applied the analyst-coverage screen row-by-row and clustered by company only. Standard errors (in parentheses) are two-way clustered by company (gvkey) and fiscal year. Firm controls are the four of Section 2 and analyst coverage enters in logs, as in Table 2; earlier versions of this table also carried leverage and entered coverage untransformed. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

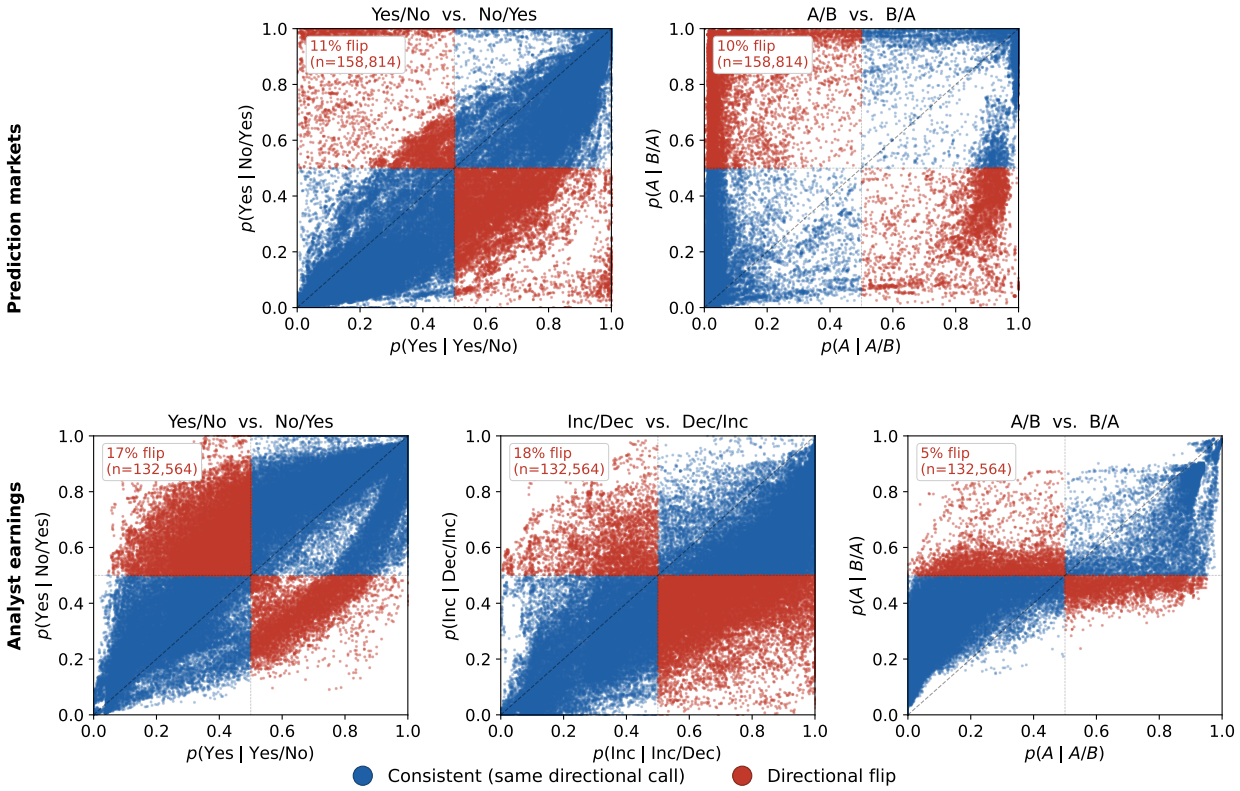
Table IA12: Directional Decomposition by Model

	LLaMA 3.1-8B	LLaMA 3.3-70B	Qwen 2.5-3B	Qwen 2.5-7B	Qwen 2.5-32B	Qwen 2.5-72B
Panel A: Prediction Markets						
Exposure	+0.0638***	−0.0010	+0.0417**	−0.0039	+0.0320	−0.0176
	(0.013)	(0.026)	(0.016)	(0.017)	(0.021)	(0.023)
Outcome (YES)	+0.5102***	+0.1167***	−0.2467***	−0.6111***	−0.2254***	−0.2051***
	(0.009)	(0.014)	(0.010)	(0.011)	(0.012)	(0.013)
Exposure × Outcome	−0.0937***	+0.1787***	+0.0421*	+0.0110	+0.0768**	+0.1823***
	(0.020)	(0.033)	(0.023)	(0.023)	(0.030)	(0.028)
Observations	158,410	158,410	120,175	120,175	120,175	120,175
Panel B: Analyst Earnings						
Exposure	+0.0292	+0.4800***	+0.0815	+0.0997	+0.3183**	+0.3527***
	(0.056)	(0.099)	(0.097)	(0.067)	(0.147)	(0.117)
Outcome (Increase)	−0.4906***	+0.0377	−0.2167***	−0.7512***	+0.4052***	+0.3101***
	(0.012)	(0.041)	(0.053)	(0.014)	(0.026)	(0.012)
Exposure × Outcome	−0.0969***	−0.3106**	+0.0028	−0.0527*	−0.2620*	−0.2862**
	(0.016)	(0.153)	(0.090)	(0.029)	(0.134)	(0.123)
Observations	132,580	132,580	130,971	130,971	130,971	130,971

Notes: Full column-(4) directional specification of Table 4, per model. Panel A is prediction markets (source FE, log volume, exposure × topic interactions, weekly calendar FE), with each model's demographic persona crowd; Panel B is analyst earnings (firm controls, exposure × GICS sector interactions, weekly calendar FE), under the informed EPS baseline. The dependent variable is the outcome-aligned signed gap; $\hat{\lambda}$, $\hat{\gamma}$, and $\hat{\theta}$ are the reference-cell exposure effect, directional prior, and exposure-by-outcome asymmetry. The LLaMA 3.3-70B column reproduces Table 4 and gives the coefficients in Figure 7. Cutoffs: LLaMA 3.x December 2023; Qwen 2.5 January–August 2024 gray zone excluded. Standard errors two-way clustered by market and evaluation date (Panel A) or company and fiscal year (Panel B). $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

Figure IA2: LLM Probabilities Under Reversed Label Orderings

Alt text: Five scatter panels, two for prediction markets above and three for analyst earnings below. Each plots the probability elicited under one label ordering against the probability under the reversed ordering, with a dashed diagonal and dividing lines at one half. Points keeping the same directional call are blue and those crossing one half are red. Each panel reports its flip rate, ranging from five percent to eighteen percent.



Notes: Each panel plots the LLM's probability under one label ordering against the reversed ordering for the same event; points off the 45-degree line move with prompt form, and red points cross the 0.5 boundary and flip the directional call. The top row is prediction markets and the bottom row analyst earnings, each shown under several equivalent label framings, with panel labels reporting the share of events whose call flips. Model: LLaMA 3.3-70B.