Cognitive Embeddings

Laurenz De Rosa

University of Chicago - Booth School of Business and Griffin Department of Economics

May 24, 2026

Preliminary and incomplete

This paper develops a theory-grounded AI model that mimics analysts' cognitive process of selecting, interpreting, and forecasting from corporate disclosures. The model uses analyst reports to identify economically relevant information in earnings calls and transforms the resulting attention-weighted call representations into horizon-specific cognitive embeddings. These embeddings are behaviorally disciplined cognitive-state proxies: model-implied representations of call content constrained by the analyst's own writing and validated against realized outcomes. On analyst-disjoint test sets, the model explains the term structure of EPS forecasts, from short- to medium- and long-term horizons, as well as long-term growth forecasts. The recovered attention measures also explain forecast errors, linking differences in analysts' written interpretations of common disclosures to differences in subsequent forecast performance. Analysts make different forecasts, and systematic forecast errors, because they attend to different parts of the same public disclosure. More broadly, the framework provides a general empirical approach for recovering model-implied information selection from behavioral text. While the paper focuses on analysts and earnings calls, the same architecture can be applied to other expectation-formation environments, including macroeconomic forecasting, retail trading, managerial communication, and investor-relations settings.

I thank Fernando Alvarez, Malcolm Baker, Leland Bybee, Lars Peter Hansen, Zhiguo He, Kilian Huber, Ralph Koijen, Eric Milstein, Sanjog Misra, Stefan Nagel, Pascal Noel, Yinan Su for an excellent discussion, Suproteem Sarkar, Harald Uhlig, Kaushik Vasudevan, Anthony Lee Zhang, and all the participants of the G3 Seminar, the Yiran Fan Memorial Conference, and the Asset Pricing Working Group for helpful comments. This research was funded in part by the John and Serena Liew Fellowship Fund at the Fama-Miller Center for Research in Finance, University of Chicago Booth School of Business. All remaining errors are my own.
First version: May 2025. De Rosa: University of Chicago (lderosa@chicagobooth.edu).

Please click here for the latest version

1 Introduction

Economic agents form expectations by representing the world selectively: choosing what to attend to, how to weight it, and what to recall. Theories of expectation formation differ in how they specify this representation. Rational inattention restricts how much information is processed (Sims, 2003); sparsity restricts which dimensions enter at all (Gabaix, 2014, 2019); diagnostic expectations distort the weights placed on recently representative states (Bordalo, Gennaioli, La Porta, and Shleifer, 2019); memory-based theories let past episodes selectively resurface (Howard and Kahana, 2002; Polyn, Norman, and Kahana, 2009; Wachter and Kahana, 2024). Each is a claim about the internal representation the agent uses, but the representation itself is unobserved.

This paper develops a framework for recovering that representation from behavioral text and for using it to discriminate between competing theories of expectation formation. The framework requires three jointly observed pieces: an *information menu*, meaning the information the agent could in principle use; a *behavioral trace*, meaning text produced after processing that menu; and an *outcome variable*, meaning a forecast, decision, or realized error that disciplines what the representation must explain. In this application, the information menu is the matched earnings call, the behavioral trace is the analyst report, and the outcome variables are forecast values and forecast errors. Analyst forecasts, matched earnings calls, and the written reports that justify those forecasts supply exactly this triple. The setting is also informative beyond analysts themselves: a substantial asset-pricing literature treats I/B/E/S forecasts as observable stand-ins for the expectations that move prices (La Porta, 1996; Bordalo, Gennaioli, La Porta, and Shleifer, 2019, 2024). Analysts are not representative investors in a literal institutional sense, but their forecasts are among the main empirical proxies through which the literature measures representative investor or market beliefs. Understanding how analysts form those forecasts, and where their errors come from, therefore matters for the expectation channel that asset-pricing research already attributes to the market.

I begin with attention because it is the most local of these mechanisms and the one most directly identified by within-call variation: holding the call fixed, which parts of it does a given analyst's report make salient, and does that selection explain the analyst's forecast and forecast error? The empirical question is whether analyst reports reveal which parts of a common call different analysts attend to, and whether a model-implied attention measure helps explain forecasts and forecast errors out of sample. I link analyst reports to matched earnings-call transcripts and estimate a theory-grounded attention model in which report snippets retrieve call snippets. The resulting attention-weighted call representation is then used to explain the term structure of EPS forecasts, long-term growth forecasts, and horizon-specific forecast errors for analysts not seen in training. Out-of-sample explanatory power for unseen analysts is economically meaningful, and learned attention beats sparsity-matched random pooling. Performance falls when analyst reports

are shuffled across test observations, showing that matched analyst text contains economically relevant information about how common disclosures are represented.

The measurement strategy treats this as an inversion problem: we observe behavior, but we care about the hidden representation or decision process that produced it (Kleinberg, Ludwig, Mullainathan, and Raghavan, 2024). In this setting, unobserved analyst attention to parts of the call produces an analyst report and a forecast. The model inverts that process. Given the information menu, behavioral trace, and outcome variable, it asks which call snippets were likely selected. The architecture therefore does not use call text alone to predict forecasts or forecast errors. It asks the analyst's own report snippets to retrieve the call snippets most consistent with that written interpretation. The report-query/call-key architecture then uses the analyst's written trace to infer which parts of the common information menu were selected.

This distinction matters for measurement. Attention is not primarily a topic-level object in this setting. Broad economic topics such as demand, margins, guidance, competition, or capital allocation can plausibly appear in almost every report and every call, so topic shares alone do not reveal what information was selectively used. The relevant object is the interaction between reasonably available information and the analyst's written justification for a forecast: given the same call, which call snippets does a particular report make salient? Sparsity is therefore not imposed by declaring that only a few topics apply. It is estimated at the snippet level from the report–call matching problem, where the model can assign zero or low weight to available disclosures even when their coarse topic labels remain broadly relevant.

This model-implied representation is the paper's *cognitive embedding*: an attention-weighted representation of call content constrained jointly by the analyst's text and by subsequent forecast outcomes. It is not directly observed, but disciplined by observed behavior and validated by out-of-sample forecast outcomes. The architecture is designed to be interpretable, not only predictive: each forecast and forecast error maps back to specific call snippets selected by the analyst's own report. The within-call attention estimates here are the first step in a broader program. Appendix A.2 shows how the same architecture nests episodic memory, diagnostic representativeness, salience, prior anchoring, sticky updating, and confirmation bias, each identified from a different empirical moment.

Literature and contribution The paper brings together three literatures that usually remain separate: economics models of expectations under limited attention, psychology and cognitive-science models of memory and internal representation, and computer-science methods for embedding and retrieval.

Economics: expectations and constrained attention In economics, the paper relates to work on analyst expectations and asset pricing, where forecast dispersion, forecast errors, and survey ex-

pectations are linked to disagreement, overreaction, and asset-market outcomes (Greenwood and Shleifer, 2014; Nagel and Xu, 2022; Bordalo, Gennaioli, Ma, and Shleifer, 2020; Bordalo, Gennaioli, La Porta, and Shleifer, 2024). More broadly, it relates to a literature emphasizing that economically important outcomes can exhibit modest explanatory power when researchers observe only coarse proxies for the information and representations agents actually use. In finance, Roll (1988) shows that even ex post stock returns are difficult to explain using systematic factors, industry information, and public news. This connection is about measurement rather than direct comparability between forecast-level prediction and return prediction: analyst reports often contain explicit forecast language, so forecast values are partly recoverable from the analyst's own text. The closer analogue to Roll's explanatory-power challenge is the forecast-error exercise, which asks whether model-implied information selection helps explain subsequent mistakes after the stated forecast has been formed. In economics more generally, measurement error and heterogeneous decision rules are central concerns (Griliches, 1986; Heckman, 2001), while recent work in econometrics and machine learning argues that richer high-dimensional data can improve prediction and measurement in environments with complex behavioral structure (Mullainathan and Spiess, 2017; Athey and Imbens, 2019). This paper applies that logic to expectation formation by using analyst reports as behavioral traces of selective information processing. It also builds on recent finance work that uses text and machine learning to recover expectations or expectation-relevant signals from unstructured language (Bybee, 2023; Bastianello, Décaire, and Günzel, 2024; Chen, 2025; Ke, 2024). The paper also relates to recent work using analyst-report text to measure economically meaningful narratives. Chen, Hwang, and Peng (2025) use analyst reports and online stock opinions to study why investors buy short-leg securities, showing that recommendation narratives often emphasize lottery-like upside. A growing finance literature adapts embedding and transformer architectures to nontraditional economic data. Gabaix, Koijen, Richmond, and Yogo (2025) learn asset and investor embeddings from portfolio holdings, treating portfolios as analogous to sentences that reveal latent asset characteristics. Clayton and Coppola (2026) use a graph transformer on security-level holdings data to learn investor and asset representations that predict trading behavior and fire-sale vulnerability. Epstein, Wang, Choi, and Pelger (2025) use attention-based latent factors for statistical arbitrage, jointly learning factor structure and trading signals. These papers show that transformer-style architectures are becoming useful tools for learning economic representations from high-dimensional financial data. My setting differs in both object and identification: analyst reports are used as behavioral queries over a common disclosure, so the learned representation is not an asset, investor, or trading embedding, but a model-implied proxy for the information an analyst selected from the call. The paper also draws on theories of rational inattention and sparse information processing, especially Sims (2003), Veldkamp (2011), Gabaix (2014), and Gabaix (2019), which interpret heterogeneous behavior as arising from con-

strained information processing rather than from differences in information alone. The empirical contribution is to estimate a high-dimensional attention rule that connects those theories to analyst forecast-error formation.

Psychology and cognitive science: memory and internal representation The paper also relates to computational cognitive science, which studies cognition as approximate inference, memory retrieval, and decision-making under resource constraints. Classic foundations include bounded rationality and constrained information processing (Simon, 1955, 1959; Kahneman, 1973; Sims, 2003; Gabaix, 2014). More recent work formalizes cognition as resource-rational, with agents allocating limited memory, attention, and computation in approximately optimal ways given environmental structure and task demands (Griffiths, Lieder, and Goodman, 2015; Gershman, 2015). The link is especially close for models of selective memory retrieval, where current goals or cues retrieve task-relevant representations from a larger memory store through similarity-based competition (Howard and Kahana, 2002; Polyn, Norman, and Kahana, 2009). It also connects to recent evidence that attention and memory constraints distort people's mental models of information environments, so behavior can reflect biased internal representations rather than preferences alone (Bohren, Hascher, Imas, Ungeheuer, and Weber, 2025). The paper is also connected to work on cognitive maps, predictive state representations, and probabilistic theories of cognition, in which internal representations are compressed summaries used to organize future-relevant information and infer latent states from noisy evidence (Stachenfeld, Botvinick, and Gershman, 2017; Tenenbaum, Kemp, Griffiths, and Goodman, 2011; Goodman and Frank, 2016).

Computer science: embeddings and retrieval Relative to text-based machine-learning approaches that treat language as a bag of predictive features or a source of sentiment, the design moves from text-as-prediction toward text-as-inversion: analyst reports act as behavioral probes that retrieve the call snippets most consistent with the analyst's written interpretation. This use of query–key attention parallels modern computational accounts of key–value retrieval, in which cue representations selectively activate relevant stored representations (Ramsauer, Schäfl, Lehner, Seidl, Widrich, Gruber, Holzleitner, Adler, Kreil, Kopp, Klambauer, Brandstetter, and Hochreiter, 2021; Gershman, Fiete, and Irie, 2025). Fixed semantic embeddings make it feasible to estimate this retrieval architecture at scale while retaining economically meaningful sparsity, validation, and falsification layers.

The paper's contribution is to combine these strands in one empirical object. Analyst-report snippets act as retrieval cues, earnings-call snippets act as candidate information items, and the learned attention weights produce a behaviorally disciplined cognitive-state proxy that is validated out of sample using forecast outcomes.

Methodological contribution Methodologically, the paper combines fixed semantic embeddings, query–key attention between analyst reports and earnings calls, analyst-disjoint out-of-sample validation, and a query-shuffle falsification to recover behaviorally disciplined cognitive-state proxies. The contribution is not a broad survey of text methods, but a structured empirical design for measuring which parts of a common disclosure different analysts appear to use when producing forecast errors. Topic labels remain useful for interpreting the recovered weights after estimation, but they are not the primitive measurement device. More broadly, the approach is flexible: any setting with a common information environment, behavioral text, and an outcome can in principle be used to recover model-implied information selection. This makes the framework adaptable to behavioral extensions such as memory over prior disclosures, selective retrieval from news, peer or consensus anchoring, macroeconomic forecasting, retail-trading narratives, managerial communication, and investor-relations language. Appendix A.2 describes these extensions and clarifies how they relate to the main within-call design.

Organization Section 2 describes forecasts, reports, calls, forecast-error construction, and sample selection. Section 3 outlines snippet construction, embedding inputs, the attention model, and the validation design. Section 4 gives a short interpretation of the attention operator as constrained information processing and clarifies scope. Section 5 reports analyst-disjoint out-of-sample performance, the query-shuffle falsification, embedding dimensionality, and the main attention-interpretation results. Section 6 concludes. The Appendix reports the behavioral extensions and complete taxonomy heatmaps for the factor-sort exercise.

2 Data and Measurement

The analysis uses three linked data sources: analyst earnings forecasts, the corresponding analyst reports, and earnings call transcripts matched to each report within a 7-day window. This matching assumes that analysts process the information from the call, form their earnings forecast, and subsequently justify it in the report.

Analyst Forecasts I/B/E/S (Institutional Brokers' Estimate System) is the standard dataset that provides detailed information on analysts' earnings forecasts, stock recommendations, and other financial estimates for publicly traded firms. The analysis focuses on earnings forecasts, specifically analysts' projections of a firm's future earnings per share (EPS) at various horizons, including the next quarter, the next fiscal year, and long-term growth over the business cycle (3–5 years). The forecast and recommendation data are obtained through WRDS I/B/E/S Detail and I/B/E/S Recommendations. The forecast data are anonymized using a numerical analyst identifier. To recover analyst identities, I link I/B/E/S Detail forecasts to I/B/E/S recommendations by matching the anonymized analyst ID *ANALYS* with *AMASKCD*, which provides the analyst's last name and first

initial. While the analyst codes are consistent across the two sources, the broker identifiers are not. To de-anonymize broker affiliations, the analysis uses the mapping provided by Law (2022), which links the anonymized forecast affiliation (*ESTIMATOR*) to its de-anonymized counterpart (*ESTIMID*).

Forecast-error construction For the annual EPS horizons, the target is the signed forecast error

$$y_{t,h} = \frac{\widehat{\mathrm{EPS}}_{t,h} - \mathrm{EPS}_{t,h}^{\mathrm{realized}}}{P_t},$$

where $\widehat{\mathrm{EPS}}_{t,h}$ is the analyst EPS forecast issued with report t, $\mathrm{EPS}_{t,h}^{\mathrm{realized}}$ is realized EPS for the same forecast horizon, and P_t is the report-date stock price used in the report-level match. Positive values therefore indicate *ex post* optimism: the analyst forecast exceeded realized EPS. The model predicts this signed error, and several descriptive splits also use its absolute value. LTG is reported separately because it is a long-term growth-rate forecast rather than a price-scaled EPS level.

Analyst Reports Analyst reports are manually collected via the LSEG Refinitiv Portal, which does not support automated API access. The current report collection contains approximately 500,000 analyst reports.[1] Each analyst in the sample is de-anonymized to the level of last name and first initial. The matching procedure begins by identifying each analyst's most frequently covered firms in the I/B/E/S dataset. The company research function on Refinitiv is then used to shortlist potential analyst matches. To verify the full identity, FINRA BrokerCheck is used to confirm first names and broker affiliations. The final match is based on the overlap of (i) firm coverage in I/B/E/S and the written reports, (ii) broker affiliation as listed on BrokerCheck, and (iii) the analyst's last name and first initial. Because this collection is manual and the paper also requires a matched call transcript and forecast target, the estimation sample is not a random draw from the I/B/E/S universe. It is tilted toward firms and analysts with richer report, transcript, and forecast coverage.

Earnings Call Transcripts (Capital IQ) Capital IQ Transcripts provide historical earnings call transcripts for approximately $8,000$ public companies globally. The data include various call types, such as earnings announcements, guidance updates, and special events, with coverage spanning multiple countries and years. Transcripts are collected and processed by S&P Global and accessed through the WRDS Capital IQ Transcripts package. The estimation sample requires a matched earnings call and forecast target, valid report and call embeddings, call sequence zero, and a report-call gap of at most seven days.

The LTG target is measured as a long-term growth-rate forecast rather than as a one-year EPS forecast scaled by price, which is why its scaled-error distribution is much more dispersed than

[1] For S&P 500 firms, the Refinitiv archive contains approximately 4.5 million analyst reports. In a future iteration, I will use all available reports for U.S. companies.

the annual EPS horizons. I therefore report LTG separately throughout the analysis and avoid interpreting its magnitude as directly comparable to the 1y–5y errors.

Before embedding, I use a language model only to segment the raw reports and call transcripts into economically meaningful snippets and to remove non-substantive boilerplate such as call formalities, operator language, report disclaimers, and rating boilerplate. The prompts explicitly instruct the model not to invent content, not to merge distinct themes, and not to neutralize tone. Thus the snippet layer is intended to preserve the original management or analyst language and slant rather than amend the source text with new LLM-generated economic judgments. Appendix A.1 reports the exact prompts.

3 Model

3.1 Snippet construction and embedding inputs

The text inputs are first converted into item-level economic snippets. A snippet is a short, self-contained semantic unit that expresses one local claim, fact pattern, or interpretation. These units are not raw pages and not fixed-length token chunks. They are closer to sentence-level or short-paragraph extracts: report snippets preserve analyst-written claims, while call snippets preserve management or analyst-call content from the matched earnings-call transcript. The extraction prompt asks the model to keep the source language and tone and to remove only non-substantive procedural or boilerplate text. It does not ask the model to score, rewrite, or reinterpret the economics of the source document. Each snippet is also assigned a short topic label, but the baseline attention model uses only the embedding vectors and not the topic labels. In the interpretation exercises below, these raw topic labels are mapped into the keyword taxonomy.

Let $\{q_{it}\}_{i=1}^{N_q(t)}$ denote report snippets (queries) and $\{k_{jt}\}_{j=1}^{N_k(t)}$ denote call snippets (keys) for analyst–firm observation t. The snippet text is embedded with OpenAI's `text-embedding-3-large` model, which produces a 3,072-dimensional vector. The empirical specification uses the first $d = 280$ coordinates of this vector, exploiting the Matryoshka property that earlier coordinates are trained to preserve most of the semantic information (Kusupati, Bhatt, Rege, Wallingford, Sinha, Ramanujan, Howard-Snyder, Chen, Kakade, Jain, and Farhadi, 2022). Each retained vector is row-normalized to unit length before estimation, so similarity is cosine similarity by construction. The use of fixed semantic embeddings builds on distributed and contextual sentence representations in NLP (Mikolov, Chen, Corrado, and Dean, 2013; Devlin, Chang, Lee, and Toutanova, 2019; Reimers and Gurevych, 2019).

The estimation sample is organized at the observation level: each analyst–firm report contains a set of report embeddings, a set of call embeddings, the report–call day gap normalized by the seven-day matching window, and scaled-by-price forecast errors for the available forecast hori-

zons. Table 1 reports the resulting snippet granularity. This scale is important for interpreting the attention diagnostics: HHI and effective keys are measures of concentration across economically meaningful call snippets, not across raw tokens, sentences, or pages.

Table 1: Snippet granularity

Object	Mean	SD	P25	Median	P75	P90
Report snippets per observation, $N_q(t)$	16.1	8.9	9	14	22	28
Call snippets per observation, $N_k(t)$	30.2	15.8	14	31	43	51
Words per report snippet	34.2	15.5	23	32	42	54
Words per call snippet	43.6	17.8	31	41	53	67

Notes: Snippet counts are computed on the full estimation sample after the report-call matching and deduplication filters. The model allows up to 128 report snippets and 256 call snippets per observation; in the estimation sample the observed maxima are 124 and 96, so these caps do not bind. Word-count moments are computed for observations whose full snippet text is retained for embedding construction. Attention concentration statistics in the paper should therefore be interpreted as concentration across roughly 30 call snippets per observation, not across raw pages or individual tokens.

3.2 Illustrative Examples

The first step in the construction is semantic matching between what the analyst writes and what management says on the call. Table 2 gives two concrete examples. The Amazon pair has cosine similarity 0.806 and illustrates economically close retrieval without near-verbatim language; the Chipotle pair has cosine similarity 0.916 and shows that near-restatements also appear in the matched text. These are the kinds of pairs the query–key layer can detect, but the attention model is not an exact-copy detector: it aggregates selected call snippets into a horizon-specific representation rather than relying only on one high-similarity match.

Table 3 gives a second example closer to the paper's identification logic. Two analysts write reports after the same IBM earnings call and make one-year EPS forecasts for the same horizon, but their forecasts and realized forecast errors differ sharply. Analyst A is optimistic ex post, with forecast minus realized EPS equal to 1.36, while Analyst B is pessimistic ex post, with forecast minus realized EPS equal to -0.28. Relative to the same-call, same-horizon forecast distribution, Analyst B is also far below the peer median, while Analyst A is close to the peer median. The recovered attention maps also differ in composition and concentration: Analyst A's top three selected call keys absorb 75.9% of the attention budget, while Analyst B's top three absorb 26.9%. One report therefore produces a concentrated selected-information set centered on management EPS targets, software acceleration, and margin commentary, while the other spreads attention more broadly across storage, semiconductor R&D, and mainframe information. The example is not a causal decomposition of the forecast gap. It shows the object the model recovers: for the same

Table 2: Illustrative semantic matches

Source	Company	Date	Original quote
Panel A. Amazon capex, cosine similarity 0.806			
Analyst report	Amazon.com (AMZN)	2023-08-04	Management now expects capex for FY23 to be slightly over $50bn, down y/y, with fulfillment and transportation investments expected to come down y/y.
Earnings call	Amazon.com (AMZN)	2023-08-03	Looking ahead to the full year 2023, we expect capital investments to be slightly more than $50 billion compared to $59 billion in 2022. We expect fulfillment and transportation CapEx to be down year-over-year, partially offset by increased infrastructure CapEx to support growth of our AWS business.
Panel B. Chipotle catering, cosine similarity 0.916			
Analyst report	Chipotle Mexican Grill (CMG)	2014-01-31	Catering contributed 1% to average check and is now 1% of sales nationally (ex. NYC, coming this year). In some markets (Denver, Seattle, St. Louis, San Francisco) catering already represents 1.5% of sales. We believe this can be a significant driver going forward.
Earnings call	Chipotle Mexican Grill (CMG)	2014-01-30	Catering sales are approaching 1% with most of this being incremental. In some of our markets, like Denver, Seattle, St. Louis and San Francisco, catering sales are averaging about 1.5% of sales. Given the strength we're seeing with our catering program so far, we're optimistic that we can continue growing our catering sales in the future.

Notes: The report snippets are matched to prior-day earnings-call snippets. The Amazon example illustrates economically close semantic retrieval without near-verbatim language; the Chipotle example shows that near-restatements also appear in the matched text, which is why the empirical model aggregates selected snippets rather than relying only on exact or near-exact matches.

public disclosure and same forecast horizon, different report language can imply different selected call information. This same-call comparison is illustrative; the systematic versions appear in Sections 5.8–5.9.

Table 3: Same call, same horizon, different selected information: IBM

	Analyst A	Analyst B
Report date	2014-07-23	2014-07-18
1y EPS forecast	17.89	10.05
Forecast error	1.36	-0.28
Effective keys	4.1	18.1
Top-3 budget	75.9%	26.9%
Top selected key 1	2014–2015 operating-EPS targets and growth-market platform commitments (32.9%)	Storage hardware weakness with flash and Storwize growth (11.4%)
Top selected key 2	Second-half software acceleration (22.6%)	Semiconductor R&D, 7-nanometer scaling, and post-silicon technologies (8.2%)
Top selected key 3	Gross-margin expansion alongside services and systems margin pressure (20.4%)	System z mainframe cycle performance (7.3%)

Notes: Both reports are matched to IBM's July 17, 2014 earnings call and use the one-year EPS forecast horizon. Forecast error is forecast minus realized EPS for the relevant report-specific one-year forecast target, so positive values are ex-post optimistic and negative values are ex-post pessimistic. Among reports matched to the same IBM call and one-year horizon, the median EPS forecast is 17.895; Analyst A is essentially at this median, while Analyst B is 7.845 EPS points below it. The table reports value-model sparsemax attention. Effective keys are $1/\sum_j w_j^2$. Top-3 budget is the sum of the three largest call-key attention weights. The two reports' top-five selected call snippets have zero overlap. The descriptions paraphrase the selected call snippets rather than quoting the transcript verbatim.

The final step is attention-weighted aggregation. For an IBM report dated April 18, 2012, the three-year cognitive embedding is approximately

$$e_{t,3y} \approx 0.487\, k_{\text{revenue growth}} + 0.212\, k_{\text{cash flow}} + 0.149\, k_{\text{competitive position}} + \cdots .$$

Here $k_{\text{revenue growth}}$, $k_{\text{cash flow}}$, and $k_{\text{competitive position}}$ denote the embeddings of the call snippets summarized by the subscripts, and the coefficients are the learned attention weights for this report–horizon pair. The largest attended call snippets emphasize revenue of $24.7 billion and growth markets outpacing major markets; $1.9 billion of free cash flow and $3.9 billion returned to shareholders; and software share gains with competitive displacements. Thus the recovered embedding is not just an average of the call. It is a point in semantic space pulled toward the dimensions the model identifies as most relevant for that analyst–horizon pair: revenue momentum, cash generation, and competitive software positioning.

Figure 1 visualizes this geometry by forming a three-dimensional basis from the top attended semantic directions. The figure projects both the full cognitive embedding and the weighted sum of the top three snippets into that span. The red star is the projected full embedding; the orange diamond is the top-three approximation. Their proximity shows that the recovered embedding is largely governed by these attended economic directions.

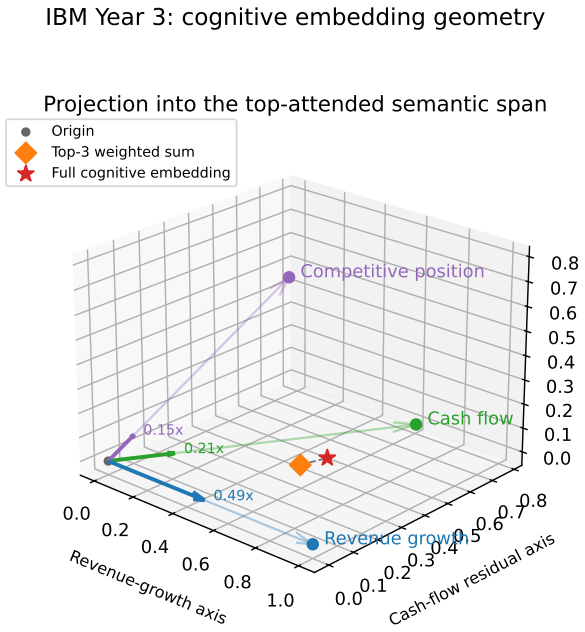


Figure 1: Projection of a cognitive embedding into the span of its top attended semantic components. The red star is the full attention-weighted cognitive embedding after projection into this three-dimensional span. The orange diamond is the approximation obtained using only the top three attended snippets. Their proximity indicates that the embedding is largely governed by the three displayed components.

This mirrors the familiar intuition behind embedding-space vector arithmetic, except here the arithmetic is attention weighted and observation specific. The model forms a cognitive embedding by adding semantic directions from the call with learned weights, so changing the attended snippets changes the economic interpretation of the recovered cognitive-state proxy.

3.3 Attention model for cognitive embeddings

For each analyst–firm report, report snippets enter as queries and matched earnings-call snippets enter as keys. The projected vectors are normalized to unit length, every report–call snippet pair is scored by cosine similarity, sparsemax is applied over call snippets, the resulting attention weights are averaged across report snippets, and the selected call snippets are pooled into a horizon-specific cognitive embedding. A fixed-form prediction head then maps that embedding into the standardized forecast error for the corresponding horizon.

Figure 2 summarizes the cognitive-transformer interpretation of this architecture. The analyst report is the observed behavioral trace, the earnings call is the common information environment, and attention recovers the call representation made salient by the report for horizon h. The figure is illustrative: the empirical object is the estimated attention-weighted call embedding $e_{t,h}$.

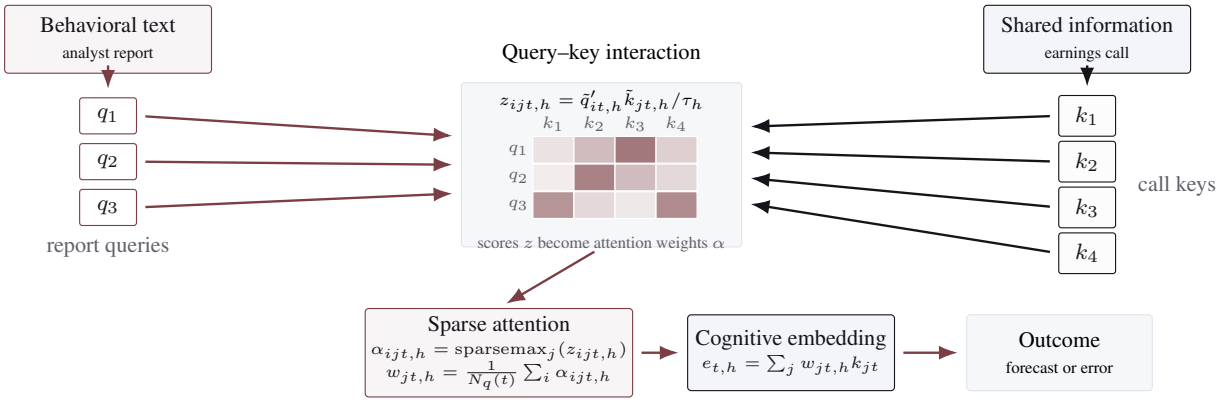


Figure 2: The cognitive-transformer architecture. Report snippets act as behavioral queries over the matched earnings call. The model scores every report–call snippet pair, converts those scores into horizon-specific attention weights over call snippets, and pools the selected call embeddings into $e_{t,h}$. The prediction head maps this selected-call representation into the forecast value or forecast error.

The baseline specification estimates a single multi-horizon attention model for six forecast targets: long-term growth and one- through five-year earnings forecasts. The outcome is the analyst's forecast error scaled by price. The model is evaluated on an analyst-disjoint split, so out-of-sample performance is measured on analysts who are not observed during training. This split is important because the object of interest is not only prediction for familiar analysts, but whether the recovered representation generalizes to unseen forecasters.

The split is performed at the analyst level rather than at the report level. Each analyst identifier is assigned to exactly one of the training, validation, or test partitions, and all reports written by that analyst follow the same assignment. The validation split is used to choose checkpoints and validation-selected tuning parameters, while the test split is held out for final reporting. This design rules out a weak form of generalization in which the model learns an analyst's stable writing style, forecast level, or vocabulary from earlier reports and is then evaluated on later reports by the same analyst. It is not a firm-time split; the goal is instead to ask whether the report–call retrieval rule learned from one set of analysts transfers to analysts whose reports never enter estimation.

Table 4 reports the main architecture and training hyperparameters. The package of modeling choices is chosen to make the recovered representation interpretable while keeping the supervised component limited. I use a fixed general-purpose embedding model rather than a finance-specific classifier such as FinBERT because the task is symmetric semantic retrieval between analyst-report

snippets and call snippets, not sentiment classification or token-level finance labeling. The $d = 280$ Matryoshka prefix is a compact representation that preserves the ordered semantic coordinates of the original embedding while reducing the size of the learned query–key maps; Section 5.5 reports the corresponding dimensionality checks. Sparsemax is the headline attention operator because exact zero weights define selected call snippets and make effective-key counts, topic inclusion, and extensive-margin exercises economically meaningful. Softmax is reported as a dense-attention predictive benchmark and often performs better, so the sparsemax choice should be read as an interpretability choice rather than a claim of predictive dominance. Finally, the one-hidden-layer MLP head with width 128 and dropout 0.1 is deliberately small and held fixed across specifications. It gives the selected-call embedding a modest nonlinear forecast map while preventing horizon-specific architecture search from becoming the source of the results.

For observation t, let $Q_t \in \mathbb{R}^{N_q(t)\times d}$ be report-snippet embeddings, $K_t \in \mathbb{R}^{N_k(t)\times d}$ be call-snippet embeddings, and $y_t \in \mathbb{R}^6$ be the six-horizon forecast-error target vector, with EPS horizons price-scaled and LTG kept in native growth-rate units. Because not every observation has every forecast horizon, the model also observes a mask $m_t \in \{0,1\}^6$. Query and key rows are normalized to unit length before estimation. Throughout this section, $h \in \{0,1,\dots,5\}$ indexes the six horizons, with $h = 0$ denoting LTG and $h = 1,\dots,5$ denoting the one- through five-year EPS horizons. The indices i and j refer to report and call snippets. The linear maps $W_{q,h}$ and $W_{k,h}$ project the d-dimensional embeddings into the horizon-specific attention space. Superscript "std" denotes standardization using training-sample means and standard deviations.

Horizon-specific query–key attention For each forecast horizon h, the model uses separate query and key projections:

$$\tilde{q}_{it,h} = W_{q,h} q_{it}, \qquad \tilde{k}_{jt,h} = W_{k,h} k_{jt}.$$

Attention logits are

$$z_{ijt,h} = \frac{\tilde{q}_{it,h}^{\top} \tilde{k}_{jt,h}}{\tau_h}, \qquad \tau_h > 0,$$

so baseline within-call attention is semantic attention. The report–call day gap is constant across call snippets within an observation, so including it in the logit would only translate every row by the same amount and would not affect the allocation of attention across snippets. Estimating within-call recency preferences would require snippet-specific timestamps.

Applying sparsemax over call snippets gives $\alpha_{ijt,h}$, and the horizon-specific call-snippet weight is

$$w_{jt,h} = \frac{1}{N_q(t)} \sum_{i=1}^{N_q(t)} \alpha_{ijt,h}.$$

For each report snippet i, observation t, and horizon h, these row-level attention weights satisfy

Table 4: Main model hyperparameters

Panel A. Architecture	
Semantic embedding dimension d	280
Attention projection dimension	128
Forecast horizons	6
Query–key projections	Horizon-specific dense maps, no bias
Attention operator	Sparsemax
Temperature initialization τ_h	1.0
Input normalization	Layer normalization, then unit-vector normalization
Prediction-head input	Selected-call embedding only
Prediction-head hidden width	128
Prediction-head activation	GELU
Prediction-head dropout	0.1
Prediction-head parameters per horizon	36,657
Maximum report snippets	128
Maximum call snippets	256
Panel B. Training	
Optimizer	AdamW
Training epochs	10
Checkpoint selection	Best analyst-disjoint validation R^2
Learning rate	5×10^{-4}
Weight decay	10^{-4}
Loss	Masked Smooth-L_1 loss
Huber threshold	1.0
Horizon weighting	Inverse-frequency square-root weights
Gradient clipping	1.0
Train/validation/test split	Analyst-disjoint
Random reference draws	20

Notes: The table reports the hyperparameters for the main sparsemax attention models used for forecast errors and forecast values. The forecast-value model uses the same architecture and training settings, changing only the supervised target. The prediction head receives $e_{t,h}$, the attention-weighted call embedding, and does not receive the report embedding or query summary directly.

$\alpha_{ijt,h} \geq 0$ and $\sum_{j=1}^{N_k(t)} \alpha_{ijt,h} = 1$. Sparsemax can set some $\alpha_{ijt,h}$ exactly to zero, which is why the resulting attention map has an interpretable selected information set. The horizon-specific cognitive embedding is the attention-weighted call embedding

$$e_{t,h} = \sum_{j=1}^{N_k(t)} w_{jt,h} k_{jt}.$$

Prediction head Each horizon has its own prediction head,

$$\hat{y}_{t,h}^{\text{std}} = g_h(e_{t,h}),$$

where g_h is a small horizon-specific neural map applied only to the selected-call embedding $e_{t,h}$. In the main specification, the head does not receive the report embedding or the query summary directly. It applies layer normalization to $e_{t,h}$, passes the result through one hidden layer with GELU activation and dropout, and then maps the hidden representation to a single standardized prediction for horizon h. The forecast-value specification uses the same fixed-form head: 280 input dimensions, hidden width 128, dropout 0.1, and a separate head for each of the six horizons. With $d = 280$ and hidden width 128, each horizon-specific head has $2d+(d\times 128+128)+(128+1) = 36{,}657$ trainable parameters, for $219{,}942$ head parameters across the six horizons. Targets are standardized using training-sample moments and mapped back to the original units for R^2 evaluation. The prediction head matters for interpretation. The cognitive embedding $e_{t,h}$ is the recovered representation; the head g_h is the supervised map from that representation to the target. Some predictive power can therefore come from the head rather than from attention alone. To keep this distinction disciplined, the head architecture is held fixed across the main attention specifications, and the projection ablations in Section 5.4 vary the query–key geometry while leaving the head size fixed. These ablations ask whether performance changes when the attention stage is constrained; they would be uninformative if the head were absorbing all economically relevant variation.

Masked sqrt-reweighted objective Let n_h be the number of nonmissing training observations for horizon h and $\bar{n}$ the average horizon coverage. Horizons are weighted by inverse-frequency square-root weights,

$$\omega_h \propto \sqrt{\frac{\bar{n}}{n_h}}, \qquad \frac{1}{6}\sum_h \omega_h = 1.$$

The model is trained with masked Smooth-L_1 loss,

$$\mathcal{L}_t = \frac{\sum_{h=0}^{5} m_{t,h}\omega_h \, \ell_{\text{Huber}}(\hat{y}_{t,h}^{\text{std}}, y_{t,h}^{\text{std}})}{\sum_{h=0}^{5} m_{t,h}\omega_h}.$$

Here ℓ_{Huber} is the Smooth-L_1 loss applied to standardized targets, and the denominator averages the loss over the horizons observed for report t. All trainable model parameters are optimized jointly with AdamW (Kingma and Ba, 2014; Loshchilov and Hutter, 2019), using learning rate 5×10^{-4}, weight decay 10^{-4}, and gradient clipping at 1.0. The effective key count

$$k_{t,h}^{\text{eff}} = \frac{1}{\sum_j w_{jt,h}^2}$$

is the inverse-Herfindahl number of selected call snippets.[2] Economically, it is the size of the model-implied information set: the number of equally weighted call snippets that would generate the same attention concentration as the learned attention vector. The statistic is not directly penalized during training. It is used to interpret the breadth of the selected-information set and to construct a sparsity-matched random benchmark.

Evaluation design The reported out-of-sample R^2 is computed on analysts unseen in training. The primary random benchmark uses the estimated effective key count for each observation and horizon, randomly selects $\text{round}(k_{t,h}^{\text{eff}})$ call snippets, mean-pools those selected snippets, and passes the random embedding through the learned prediction head without retraining. This benchmark holds both the approximate amount of attended text and the decoder fixed while removing the learned semantic selection rule. I also report a retrained random benchmark, where horizon-specific prediction heads are fitted on the random embeddings; this is useful as a capacity comparison but is less direct as an attention falsification.

I also report several call-only benchmarks that remove query-conditioned attention altogether. A mean-pooling benchmark averages all call-snippet embeddings and either passes the resulting embedding through the learned prediction head or fits a separate flexible predictor. A TF-IDF benchmark fits ridge regressions on hashed call-text features. A Loughran–McDonald benchmark fits ridge regressions on standard financial dictionary category shares. A topic-share benchmark fits ridge regressions on the distribution of call-snippet topic labels. These comparisons ask whether the attention model improves on flat semantic averages, lexical representations, dictionary sentiment, and low-dimensional topic summaries using the same analyst-disjoint split.

Inversion interpretation The design is motivated by the inversion problem: the representation that guides behavior is not directly observed, so it must be inferred from behavioral traces. The supervised attention model uses report snippets, call snippets, and forecast outcomes to estimate a model-implied cognitive embedding: the attention-weighted call representation that explains

[2] The square turns the attention weights into a concentration measure. If attention is split equally across m snippets, then $\sum_j w_j^2 = m(1/m)^2 = 1/m$, so $1/\sum_j w_j^2 = m$. If all attention is placed on one snippet, the same formula gives one effective key. Intermediate allocations therefore translate uneven attention weights into the number of equally weighted snippets that would generate the same concentration.

forecasts and forecast errors subject to the attention architecture. The sparsity-matched random comparison holds the amount of attended text approximately fixed, so an advantage for learned attention indicates that the embedding captures structured semantic selection rather than arbitrary sparse pooling.

4 Interpretation and Scope

The attention operator admits a familiar interpretation from theories of constrained information processing. For a given report snippet i, call snippet j, and horizon h, the empirical model forms a projected semantic score $s_{ijt,h} = \tilde{q}_{it,h}^{\top}\tilde{k}_{jt,h}$ and scales it by the horizon-specific temperature τ_h. Attention then solves a regularized allocation problem over call snippets. With softmax, the regularizer is Shannon entropy and the rule has a Sims-style dense-attention interpretation (Sims, 2003). With sparsemax, the regularizer is quadratic and the rule has a Gabaix-style sparse-attention interpretation that sets some weights exactly to zero (Gabaix, 2014, 2019). The baseline uses sparsemax not because dense attention is uniformly worse, but because exact zero weights make the selected information set economically interpretable.

The temperature parameter τ_h controls semantic selectivity after the projected queries and keys have been normalized to unit length. Lower τ_h produces sharper retrieval from the same call; higher τ_h produces broader aggregation across related snippets. Because the unit-norm projections remove the scale indeterminacy between the projection matrices and τ_h, the estimated temperature profile has the behavioral meaning used throughout the paper: attention breadth rather than arbitrary rescaling.

The scope of the interpretation is deliberately narrow. The matched-call design identifies within-call selective attention: which parts of a common disclosure different analysts appear to use when forming forecast errors. The paper does not estimate a broader memory model over prior calls, past reports, or news. Appendix A.2 clarifies how the same architecture would generalize to memory and other expectation-formation mechanisms, but the main estimates should be read as validated within-call cognitive-state proxies for forecast errors, not as complete measurements of belief formation.

5 Results

The results proceed from prediction to identification and interpretation. I first ask whether the architecture can recover analysts' forecast values themselves; this easier level-prediction exercise gives a useful benchmark for what is broadly encoded in the text environment. I then turn to the paper's main object: forecast-error prediction and validation, where the query-shuffle falsification disciplines the cognitive-state interpretation. The remaining subsections examine the embedding

representation, interpret the horizon-specific forecast-value embedding that directly recovers analysts' stated forecasts, compare value-trained attention across analyst-precision groups, and benchmark the resulting mental-model measures across standard characteristic sorts. Appendix results report behavioral extensions and complete taxonomy heatmaps for the factor-sort exercise; the Supplemental Appendix reports prompts, forecast-value transfer checks, and additional robustness evidence.

5.1 Forecasting Forecast Levels Directly

The paper's main object is forecast-error formation: which parts of a common disclosure help explain why an analyst is later too optimistic or too pessimistic. Before turning to forecast errors, a useful diagnostic is to ask whether the same architecture can recover the forecast itself. I therefore re-estimate the multi-horizon attention model with the raw forecast value as the target, using the same analyst-disjoint split, horizons, embedding depth, sparsemax attention rule, horizon reweighting, and estimation protocol as in the forecast-error specification. Formally, let $a(t)$ denote the analyst attached to observation t, and let $\mathcal{A}_{\text{train}}$, $\mathcal{A}_{\text{val}}$, and $\mathcal{A}_{\text{test}}$ be mutually exclusive analyst sets. In this subsection the supervised target is the forecast level $V_{t,h}$. The value-trained model estimates

$$\theta^V = \arg\min_{\theta} \sum_{t:\, a(t)\in\mathcal{A}_{\text{train}}} \frac{\sum_{h=0}^{5} m_{t,h}\omega_h\, \ell_{\text{Huber}}\big(g^V_{h,\theta}(e^V_{t,h,\theta}), V^{\text{std}}_{t,h}\big)}{\sum_{h=0}^{5} m_{t,h}\omega_h}, \qquad e^V_{t,h,\theta} = \sum_j w^V_{jt,h,\theta} k_{jt}.$$

Here θ^V collects all value-model parameters, $g^V_{h,\theta}$ is the horizon-specific prediction head, $w^V_{jt,h,\theta}$ are the value-trained attention weights over call snippets, and $e^V_{t,h,\theta}$ is the resulting value-trained cognitive embedding. The mask $m_{t,h}$, horizon weight ω_h, call embedding k_{jt}, and Huber loss are the same objects defined in Section 3; $V^{\text{std}}_{t,h}$ is the forecast value standardized using training-sample moments. The validation-analyst set is used for model selection, and all reported R^2 values are computed only on $\mathcal{A}_{\text{test}}$, so no analyst in the test set appears during training.

Figure 3 reports the sparsemax forecast-value and forecast-error prediction exercises side by side. The forecast-value exercise trains the report-query/call-key model on forecast levels, while the forecast-error exercise trains the same architecture on signed forecast errors. In each panel, learned sparsemax attention is compared with a same-k_{eff} frozen-random reference with the prediction map held fixed. Sparsemax gives the cleanest selection test because exact zeros define a selected information set. The corresponding softmax figure, which uses broader dense attention, is reported in Appendix A.5.

The forecast-value benchmark horserace is reported in Appendix Figure A.1. That exercise uses forecast values, not forecast errors, as the target. The sparsemax and softmax attention bars

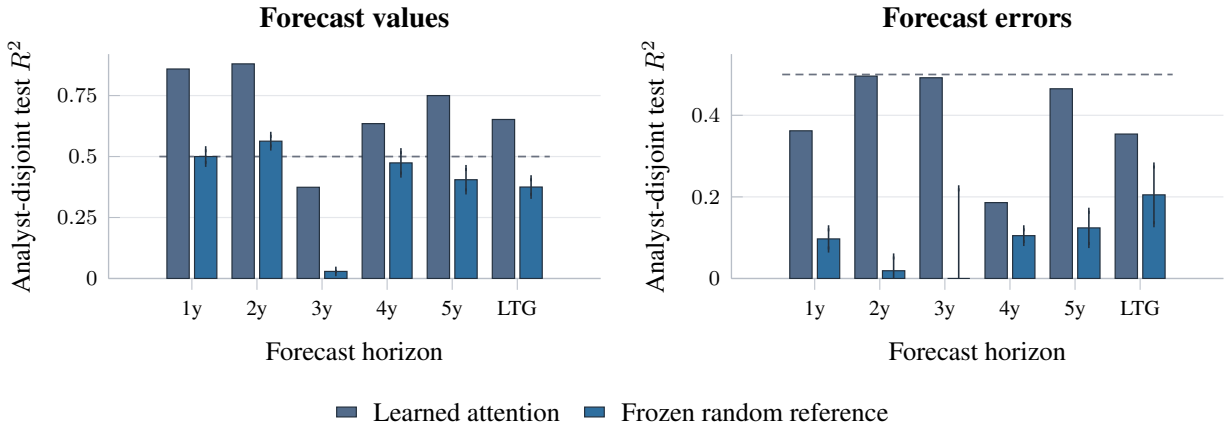


Figure 3: Sparsemax attention versus same-k_{eff} frozen-random references. The left panel reports analyst-disjoint test-set R^2 for the sparsemax attention model trained on forecast values. The right panel reports the same exercise for signed forecast errors. Frozen-random references select the same effective number of call snippets as the corresponding attention model and pass the resulting embedding through the trained prediction head with the head held fixed. Error bars show the standard deviation across random draws.

come from newly trained value-target report-query/call-key models, and the frozen-random same-k_{eff} references use the corresponding value-trained heads with the head held fixed. The mean-embedding, TF–IDF, topic-share, and Loughran–McDonald benchmarks are also re-estimated with forecast value as the left-hand-side variable on the same analyst-disjoint split.

The plotted benchmark labels are as follows. "Sparsemax attention" and "Softmax attention" are the full report-query/call-key models. "Sparsemax frozen random" and "Softmax frozen random" randomly select the same effective number of call snippets as the corresponding attention model and pass the resulting embedding through that model's value head with the head fixed. "Mean emb. frozen head" passes the mean-pooled call embedding through the sparsemax value head with estimated parameters held fixed, while "Mean emb. GB" fits a separate gradient-boosted predictor on the mean-pooled call embedding. The TF–IDF, topic-share, and Loughran–McDonald bars are ridge regressions on hashed call text, call-topic shares, and financial-dictionary category shares, respectively.

Table 6 reports the forecast-value counterparts to the attention-sparsity diagnostics, with softmax and sparsemax shown as separate panels.

The effective-key statistic gives the attention weights an economic interpretation. It is the inverse-Herfindahl size of the selected information set: the number of equally weighted call snippets that would generate the same concentration as the learned attention vector. Since the typical call contains roughly 30 candidate snippets, sparsemax effective-key counts of about 7–10 mean that the model explains forecasts using roughly one-quarter to one-third of the available disclosure, not the full call. The point is not that analysts literally hear only these snippets. Rather, the model

Table 5: Query-shuffle falsifications for forecast values and forecast errors. The exercise randomly replaces the analyst report query with a report query from another test observation, holding the earnings call and outcome fixed. Panel A uses forecast values as the outcome. Panel B repeats the same exercise for signed forecast errors.

Horizon	Correct query	Shuffled query	Difference
Panel A: Forecast values			
YR 1	0.859	0.582	0.277
YR 2	0.880	0.568	0.313
YR 3	0.374	0.040	0.334
YR 4	0.635	0.401	0.234
YR 5	0.750	0.476	0.274
LTG	0.652	0.546	0.105
Panel B: Forecast errors			
YR 1	0.362	0.140	0.222
YR 2	0.496	0.299	0.197
YR 3	0.492	0.007	0.485
YR 4	0.186	0.174	0.011
YR 5	0.465	0.372	0.093
LTG	0.354	0.304	0.050

Notes: The table reports analyst-disjoint test-set R^2 values for the full sparsemax report-query/call-key model. The shuffle randomly replaces the matched analyst report query with a report query from another test observation, holding the earnings-call snippets and the outcome fixed. Panel A uses forecast values as the target. Panel B repeats the same exercise for signed forecast errors. Differences are correct-query R^2 minus shuffled-query R^2.

Table 6: Forecast-value attention performance and concentration on the analyst out-of-sample test set. Panel A reports the value-trained softmax model and Panel B reports the value-trained sparsemax model. Attention R^2 is the full report-query/call-key model. Frozen rand. R^2 is the same-k_{eff} random-attention reference with the prediction head held fixed. *Available call snippets* is the mean number of candidate call snippets. *Effective keys* $= 1/\sum_j w_j^2$ is the effective number of attended call snippets, and *Effective keys %* scales this by the mean number of available call snippets.

Panel A: Softmax attention

Forecast	Attention R^2	Frozen rand. R^2	Avail. call snippets	Effective keys	Effective keys %
YR 1	0.873	0.332	26.79	14.39	53.72
YR 2	0.882	0.497	26.64	15.59	58.53
YR 3	0.602	-0.234	30.55	17.12	56.05
YR 4	0.738	0.515	34.25	17.58	51.33
YR 5	0.788	0.523	38.33	16.75	43.69
LTG	0.574	0.514	27.87	21.09	75.68

Panel B: Sparsemax attention

Forecast	Attention R^2	Frozen rand. R^2	Avail. call snippets	Effective keys	Effective keys %
YR 1	0.859	0.359	26.79	9.25	34.53
YR 2	0.880	0.365	26.64	9.12	34.23
YR 3	0.374	0.040	30.55	10.30	33.72
YR 4	0.635	0.317	34.25	7.00	20.45
YR 5	0.750	0.248	38.33	6.28	16.38
LTG	0.652	0.442	27.87	9.35	33.55

rationalizes their forecasts and forecast errors with a sparse selected-information set. This is the empirical analogue of constrained information processing: from a large public disclosure, only a small number of semantic directions are needed to explain the analyst-specific forecast outcome.

The forecast-level exercise performs strongly: the attention model reaches a mean analyst-OOS R^2 of 0.69, with especially high values at the one- and two-year horizons. The attention model beats sparsity-matched random pooling at every horizon, and it also improves on the mean-embedding fixed-head diagnostic at every horizon. Thus the model is not merely using a flexible head to map generic call averages into forecast levels; query-conditioned semantic selection adds information even when the target is the analyst's forecast level.

The forecast-value robustness exercises also make clear why these results are interpreted differently from the forecast-error tests. Softmax improves the mean forecast-value R^2 from 0.69 to 0.74, but uses substantially broader attention, with average effective keys of 17.1 versus 8.7 under sparsemax. This contrast is informative. Softmax uses broader attention and sometimes improves level prediction, but sparsemax achieves similar forecast-value performance with substantially fewer effective keys. Thus the predictive signal is not only present in broad call averages; it can be compressed into a smaller selected information set. This compression is what makes the attention map economically interpretable. Flat call-content benchmarks are also powerful for forecast levels: TF–IDF is competitive at several horizons, and mean-pooled call embeddings with

a gradient-boosted head beat attention for LTG and the three-year horizon. Those facts are expected if forecast levels encode persistent firm scale and information that is broadly present in the call. The value exercise therefore supports the measurement pipeline, but it does not replace the forecast-error identification tests.

Panel A of the query-shuffle exercise gives the same message from a different angle. The correct report query outperforms a shuffled report query at every forecast-value horizon, with drops of 0.28 at 1y, 0.31 at 2y, and 0.33 at 3y. Thus report queries matter even for forecast levels. At the same time, shuffled queries retain substantial explanatory power for several horizons, consistent with the idea that forecast levels contain a large common call- and firm-level component.

This check also clarifies the interpretation of the main forecast-error results. The Supplemental Appendix transfer check in Section S.13 shows that forecast-value structure transfers only partially to forecast errors: when the value-trained attention map is held fixed and passed through the saved forecast-error head without retraining, the mean analyst-OOS R^2 on scaled forecast errors is 0.191. The forecast-error-trained attention model remains stronger at most horizons. Hence the paper's main cognitive-state proxy is not simply a model for forecast levels. Forecast levels are recoverable from the text environment, but forecast mistakes require additional target-specific attention structure.

5.2 Forecast-Error Prediction and Validation

The forecast-error section repeats the same estimation design with the paper's main behavioral outcome. For annual EPS horizons, I denote the signed price-scaled forecast error by $E_{t,h} \equiv y_{t,h}$:

$$E_{t,h} = \frac{\widehat{\mathrm{EPS}}_{t,h} - \mathrm{EPS}_{t,h}^{\mathrm{realized}}}{P_t},$$

with LTG kept in its native long-term-growth units. Holding the analyst split, matched report–call inputs, horizon masks, and horizon weights fixed, the error-trained model estimates

$$\theta^E = \arg\min_{\theta} \sum_{t:\, a(t)\in\mathcal{A}_{\mathrm{train}}} \frac{\sum_{h=0}^{5} m_{t,h}\omega_h\, \ell_{\mathrm{Huber}}\big(g_{h,\theta}^E(e_{t,h,\theta}^E), E_{t,h}^{\mathrm{std}}\big)}{\sum_{h=0}^{5} m_{t,h}\omega_h}, \qquad e_{t,h,\theta}^E = \sum_j w_{jt,h,\theta}^E k_{jt}.$$

Here the superscript E marks the error-trained version of the same objects: the parameter vector, prediction head, attention weights, and cognitive embedding. The variable $E_{t,h}^{\mathrm{std}}$ is the standardized signed forecast error. As above, model selection uses only $\mathcal{A}_{\mathrm{val}}$, and all reported R^2 values are computed on held-out analysts in $\mathcal{A}_{\mathrm{test}}$. Thus the value and error exercises differ in the left-hand-side variable, not in the training, validation, and test partition.

The right panel of Figure 3 reports the corresponding sparsemax forecast-error validation. The

learned attention model improves on its same-k_{eff} frozen-random reference at every horizon, so the forecast-error signal is not explained by selecting an arbitrary sparse subset of call snippets and passing it through the trained head. The softmax counterpart is reported in Appendix A.5.

The supplement broadens the forecast-error benchmark set beyond sparsity-matched random pooling. The comparison moves from full query-conditioned semantic selectors to increasingly coarse call-only summaries, including mean embeddings, TF–IDF, dictionary categories, and topic shares; the supplemental table reports the corresponding numerical values.

The attention model clearly improves on arbitrary sparse selection, dictionary sentiment, and most topic-share baselines. However, the benchmark comparison also shows that flat call content is informative: mean-pooled embeddings and TF-IDF features are competitive at some horizons. I therefore interpret the evidence as support for a more specific claim. Query-conditioned attention is not needed because call text has no signal; it is useful because it recovers an interpretable selection rule that usually improves on arbitrary sparse selection and simple text summaries, while preserving a direct map from attended disclosures to forecast errors.

Panel B of Table 5 reports the same query-shuffle falsification for forecast errors. The exercise randomly replaces the analyst's report snippets with report snippets from another test observation while holding the earnings call and realized forecast error fixed. If the model were merely extracting generic call-level predictability, this shuffle should leave performance largely unchanged. Instead, performance falls at every horizon, with especially large drops at the one-, two-, and three-year horizons. The horizon pattern therefore supports the main interpretation: matched analyst language matters for explaining forecast errors because it changes which parts of the common disclosure are retrieved and summarized by the attention model.

5.3 Separating Selected Information from Analyst Interpretation

Section 5.2 showed that matched analyst reports contain economically relevant information beyond generic call content. This subsection asks a narrower question: conditional on the call information already selected by the attention model, does the report itself contain additional target-relevant structure?

The attention model produces the horizon-specific selected-call embedding

$$e_{t,h} = \sum_j w_{jt,h} k_{jt},$$

which summarizes the call information selected for report t and horizon h. I interpret $e_{t,h}$ as the model-implied selected-information channel. The question in this subsection is whether analyst language contains additional predictive structure beyond that selected-information channel.

To study this, I construct two report-side summaries. The first is the mean report embedding,

$$\bar{q}_t = \frac{1}{N_q(t)} \sum_i q_{it}.$$

The second removes the first-order linear overlap between the report embedding and the selected-call embedding:

$$r_{t,h}^{\text{proj}} = \bar{q}_t - \frac{\bar{q}_t' e_{t,h}}{e_{t,h}' e_{t,h}} e_{t,h}.$$

This projection residual should be interpreted narrowly. It removes only the component of the report embedding aligned with the selected-call embedding direction. It does not remove all call-related semantic information, nor does it imply a unique decomposition of embedding space into "information" and "interpretation." Instead, it provides a diagnostic measure of report-side variation orthogonal to the selected-call representation.

The baseline attention model is held fixed throughout the exercise. The saved sparsemax attention weights, the selected-call embedding $e_{t,h}$, and the baseline attention prediction

$$\hat{y}_{t,h}^A = g_h(e_{t,h})$$

are not re-estimated. I then ask whether report-side embeddings explain residual variation beyond the selected-call channel by estimating additive report corrections:

$$\hat{y}_{t,h}^G = \hat{y}_{t,h}^A + \lambda_h \phi_h(x_{t,h}),$$

where $x_{t,h}$ is either the mean report embedding or the projection residual, ϕ_h is the additive report-correction map, and λ_h is selected on the analyst-disjoint validation split. Setting $\lambda_h = 0$ exactly recovers the baseline attention model.

Table 7 reveals a sharp distinction between forecast levels and forecast errors. For forecast values, report embeddings are highly predictive even without selected-call information. This is expected because analyst reports often contain explicit forecast language, narrative framing, and persistent firm-level information directly related to stated forecasts.

For forecast errors, the selected-call embedding already captures most of the predictive signal. Mean report embeddings produce only modest incremental improvements at most horizons, suggesting that much of the forecast-error structure operates through selective interpretation of common disclosures rather than through report-side language alone. The projection residual adds more strongly at the two- and five-year horizons. This pattern suggests that medium- and long-horizon forecast errors depend more heavily on extrapolative interpretation that is not fully captured by the

Table 7: Do analyst-report embeddings add beyond selected call information?

Panel A: Forecast errors						
Horizon	Attention R^2	Mean-query only R^2	Attention + mean query R^2	Mean-query ΔR^2	Attention + projection R^2	Projection ΔR^2
YR 1	36.1	34.9	35.6	-0.5	36.1	0.0
YR 2	43.4	55.2	44.8	1.4	54.7	11.3
YR 3	49.2	63.2	49.7	0.5	49.5	0.3
YR 4	14.3	19.7	19.4	5.1	17.0	2.7
YR 5	41.7	46.1	43.4	1.7	51.2	9.5
LTG	35.6	26.9	35.6	0.0	35.6	0.0
Panel B: Forecast values						
Horizon	Attention R^2	Mean-query only R^2	Attention + mean query R^2	Mean-query ΔR^2	Attention + projection R^2	Projection ΔR^2
YR 1	85.9	77.3	88.4	2.5	81.6	-4.3
YR 2	88.0	78.3	90.7	2.6	90.0	2.0
YR 3	37.4	78.6	53.3	15.9	41.2	3.8
YR 4	63.5	66.5	75.1	11.6	74.3	10.8
YR 5	75.0	66.7	77.3	2.3	78.9	3.9
LTG	65.2	61.1	72.6	7.4	71.4	6.3

Notes: Values are analyst-disjoint test-set R^2 in percentage points. The baseline is the saved sparsemax attention model for the corresponding target. To keep the comparison aligned with the main attention architecture, the mean-query-only column fits the same MLP-head family used after the cognitive embedding to the mean analyst-report snippet embedding $\bar{q}_t$. The mean-query and projection add-ons fit MLP-head corrections to the attention-model residual, using either $\bar{q}_t$ or the residual from projecting $\bar{q}_t$ away from the attention-weighted selected-call embedding $e_{t,h}$. Add-on corrections are validation-gated, with zero gate recovering the attention baseline.

selected-call channel itself.

The results should therefore not be interpreted as a structural separation between "information" and "interpretation." Rather, they show that the selected-call embedding captures a substantial share of forecast-error-relevant structure, while residual report-side geometry contains additional horizon-dependent information.

5.4 Projection Ablation

The main attention specification uses separate dense query and key projections before computing cosine similarity. Let $W_{q,h}$ denote the report/query projection and $W_{k,h}$ denote the call/key projection for horizon h. The Appendix asks whether that projection flexibility matters for forecast-error prediction by comparing five assumptions on these matrices:

1. *Full separate*: $W_{q,h}$ and $W_{k,h}$ are unrestricted dense matrices, learned separately.
2. *Full shared*: $W_{q,h} = W_{k,h} = W_h$, so reports and calls use the same unrestricted dense matrix.
3. *Diagonal separate*: $W_{q,h} = \mathrm{diag}(a_{q,h})$ and $W_{k,h} = \mathrm{diag}(a_{k,h})$, so each side can only rescale embedding coordinates.
4. *Diagonal shared*: $W_{q,h} = W_{k,h} = \mathrm{diag}(a_h)$, so both sides use the same coordinate rescaling.

5. *Identity*: $W_{q,h} = W_{k,h} = I$, so attention is computed directly in the original normalized embedding space.

Table 8 reports the single-prediction-equivalent parameter counts for these projection-constraint families. The MLP-head column refers to the fixed neural prediction head used after attention pooling, not to the gradient-boosted benchmark in the horserace. Appendix evidence shows that the learned query–key geometry is part of the forecast-error signal rather than a dispensable parameterization: restricting the projection stage generally weakens analyst-disjoint performance, even though the prediction-head architecture is held fixed. Figure 4 repeats the same constraint exercise with forecast values as the left-hand-side variable on the full analyst-disjoint sample. In contrast to the forecast-error target, forecast-value prediction is less dependent on fully separate dense query and key projections: the full-shared and diagonal-separate restrictions essentially match the best mean out-of-sample R^2, while the identity model remains competitive. The main takeaway is that the identity-matrix models perform comparably enough to make a raw-similarity retrieval rule empirically plausible: in these specifications, the attention stage can be based directly on cosine similarity in the original normalized embedding space rather than on a learned projection geometry. The Appendix reports the corresponding parameter-count formulas.

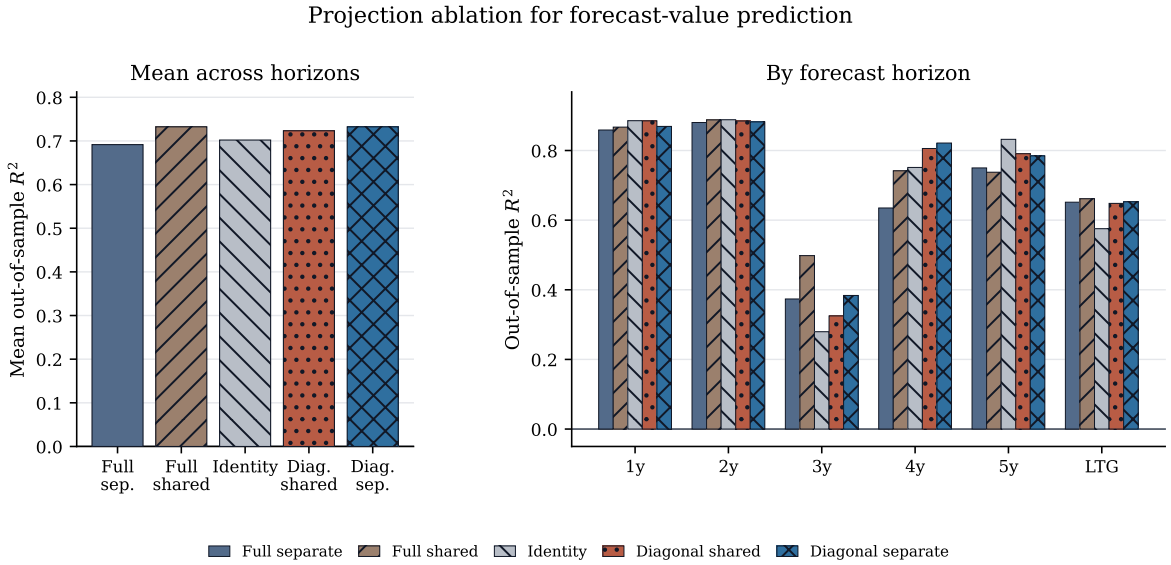


Figure 4: Projection-ablation performance for forecast-value prediction. The figure reports five projection-constraint families, now using forecast values rather than forecast errors as the outcome. The LTG column is horizon index 0, a long-term-growth forecast rather than a fiscal-year-zero forecast.

Table 8: Single-prediction-equivalent parameter breakdown for the projection-constraint families. Each row shows what one forecast-specific model would contain if the same architecture predicted only a single horizon rather than six jointly. With $d = 280$ and $d_a = 128$, one dense query or key projection contains $280 \times 128 = 35{,}840$ parameters, one diagonal scaling contains 280 parameters, and $\tau = 1$ because a single forecast has one learned temperature.

Model	Q-side	K-side	Shared	τ	Attention total	MLP head
Full separate	35,840	35,840	0	1	71,681	36,657
Full shared	0	0	35,840	1	35,841	36,657
Diagonal separate	280	280	0	1	561	36,657
Diagonal shared	0	0	280	1	281	36,657
Identity	0	0	0	1	1	36,657

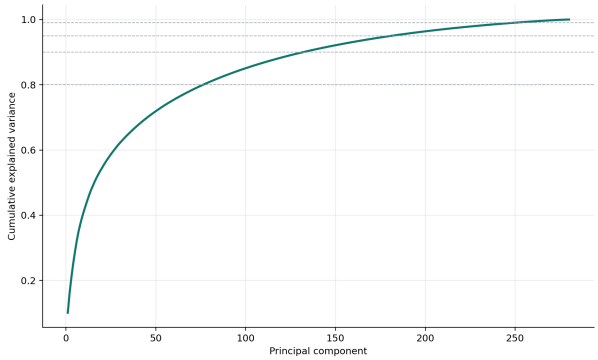
Notes. Q-side = learned query-only transform parameters. K-side = learned key-only transform parameters. Shared = one transform reused for both Q and K. τ = one temperature for the single forecast. Attention total excludes both the query/key LayerNorm parameters and the final prediction head. MLP head is the neural prediction head used after attention pooling; it is not the gradient-boosted benchmark.

Model assumptions. Full separate: Separate dense matrices for queries and keys: W_q and W_k are both learned, with no equality constraint. Full shared: One dense matrix shared by queries and keys: $W_q = W_k = W$. Diagonal separate: Separate diagonal matrices for queries and keys: $W_q = \mathrm{diag}(a_q)$, $W_k = \mathrm{diag}(a_k)$. Each feature is only rescaled, not mixed. Diagonal shared: One diagonal matrix shared by queries and keys: $W_q = W_k = \mathrm{diag}(a)$. Each feature is only rescaled, not mixed. Identity: No learned query/key projection: $W_q = W_k = I$, so attention is computed in the original normalized embedding space.

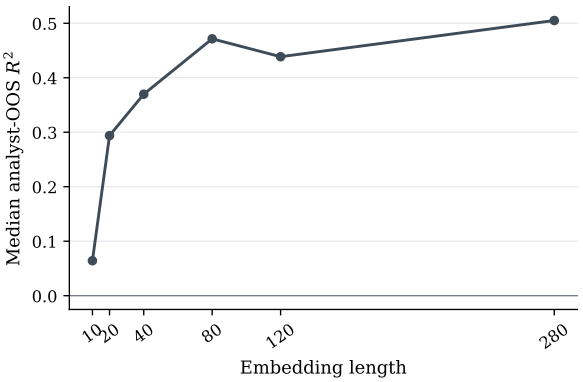
5.5 Embedding Dimensionality

The validation results establish that learned attention predicts forecast errors better than several call-level baselines. This subsection asks a narrower interpretability question: where, inside the learned cognitive embedding, does that signal live? The exercise is not meant to assign economic labels to individual coordinates. A single coordinate can mix several economic directions. Instead, I ask whether increasingly long prefixes of the Matryoshka embedding recover recognizable targets out of sample. If the early coordinates were merely arbitrary compression artifacts, they should not systematically recover the paper's attention distinctions or external firm exposures.

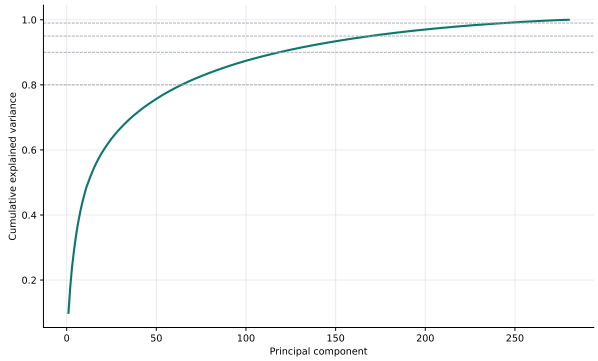
Figure 5 collects the dimensionality checks for both target specifications. A principal-component analysis of the pooled 280-dimensional cognitive embeddings gives the first check. In an unsupervised variance sense, the forecast-error representation is not low-dimensional: in the combined train-plus-test sample, 77 principal components are needed to explain 80% of variance, 132 to explain 90%, and 182 to explain 95%, with an effective rank of roughly 39. The forecast-value representation is similar but slightly more concentrated: 65, 119, and 170 components are needed to explain 80%, 90%, and 95% of variance, respectively, with an effective rank of roughly 32. The embedding space therefore does not collapse onto a few global factors in either target specification. At the same time, Figures 5b and 5d show that predictive performance rises quickly when



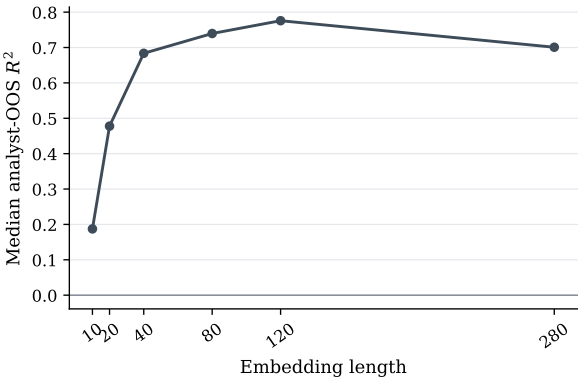
(a) Forecast-error PCA variance



(b) Forecast-error performance by length



(c) Forecast-value PCA variance



(d) Forecast-value performance by length

Figure 5: Dimensionality and predictive content of the cognitive embeddings. Panels A and B use the forecast-error attention model. Panels C and D repeat the exercise using the forecast-value attention model. Panels B and D use the same embedding-length grid and analyst-disjoint evaluation design.

the first Matryoshka coordinates are retained and then flattens. The two facts are not contradictory. Variance is spread across many directions, while forecast-relevant information is deliberately front-loaded by the Matryoshka objective.

Table 9 asks what economic information is concentrated in those early coordinate blocks. The exercise should be interpreted as a cumulative-prefix probe. Because the embedding is trained with the Matryoshka objective, the first 10, 20, and 40 coordinates are meaningful nested representations. For each table row r, I construct a target $Y_i^{(r)}$ for report observation i. For each truncation length m, I then estimate

$$Y_i^{(r)} = \gamma_{rm} + \sum_{\ell=1}^{m} \beta_{\ell rm} e_{i\ell} + u_{irm},$$

where $e_{i\ell}$ is coordinate ℓ of the report-level cognitive embedding. The right-hand side contains only the first m embedding coordinates, standardized using the training sample. The ridge estimator is fit on training reports with the same fixed penalty across rows and prefixes. The fitted model is then evaluated on held-out reports, and the table reports the resulting out-of-sample R^2.

The left-hand side changes by row. For narrative rows, $Y_i^{(r)}$ is a model-implied attention share or attention contrast. The category groups are defined as follows. "Operating evidence" is meant to capture realized business fundamentals. "Conditional information" captures disclosures whose interpretation depends more heavily on contingencies, financing conditions, management plans, or unresolved risks.

Group	Symbol	Included call-snippet categories
Operating evidence	$\mathcal{O}_i$	Segment performance; competition/pricing; customer usage; demand/revenue; margins, costs, and profitability.
Conditional information	$\mathcal{C}_i$	Guidance/outlook/expectations; operations, supply chain, and capacity; capital allocation and balance sheet; credit/funding; M&A; regulation, legal, and tax.
Demand/revenue	$\mathcal{D}_i$	Demand, revenue, orders, and growth.
Guidance	$\mathcal{G}_i$	Guidance & Outlook; Guidance & Expectations.
Balance/capital	$\mathcal{B}_i$	Capital allocation, cash, balance sheet, credit, lending, and funding.
Commercial/product	$\mathcal{P}_i$	Demand/revenue, product and technology, customer usage, customer engagement, and commercial quality.

The operating-minus-conditional row is high when the model attends more to realized business fundamentals than to caveated or contingent information. Each attention-share target is constructed by summing the attention weights on snippets in the corresponding group; each contrast subtracts one such group sum from another. These variables are attention allocations, not assessments of tone. For example, guidance attention measures how much attention is placed on guidance snippets, not whether the guidance is positive or negative. The narrative rows are therefore internal interpretability probes: they ask whether the embedding prefix retains the model-implied allocation of attention across economically named call content. The non-narrative rows are external probes. For each covered firm, I estimate trailing factor betas from monthly excess returns over

the prior 36 months, requiring at least 24 monthly observations. The displayed factor probes are market, HML (value minus growth), UMD (momentum), RMW (profitability), and CMA (conservative investment), matching the factor set emphasized below. Each cell is the held-out R^2 from a separate regression. The entries are fit measures, not attention shares, so they do not add up to one; a negative value means the truncated embedding predicts worse than the held-out mean.

Table 9: Predictive content of early embedding prefixes. Each column truncates the 280-dimensional cognitive embedding to its first m Matryoshka coordinates and fits the same ridge probe from that cumulative coordinate block to the row target. Narrative rows use model-implied attention shares or contrasts; factor rows use trailing monthly factor betas. Entries are out-of-sample R^2 values on held-out reports. Larger values mean that the target is more recoverable from the retained prefix; negative values mean worse fit than the held-out mean.

Target	10 dims	20 dims	40 dims	80 dims	160 dims	280 dims
Narrative attention targets						
Operating evidence attention	0.01	0.10	0.23	0.32	0.37	0.43
Conditional information attention	0.04	0.12	0.15	0.18	0.20	0.25
Operating evidence − conditional information	-0.02	0.10	0.21	0.29	0.35	0.42
Demand/revenue attention	0.06	0.20	0.33	0.41	0.44	0.49
Guidance attention	0.04	0.06	0.08	0.10	0.14	0.17
Balance/capital vs. commercial/product	-0.02	0.10	0.20	0.29	0.37	0.43
Return-factor exposures						
Market beta	0.01	0.03	0.01	0.04	0.09	0.15
HML beta	0.06	0.12	0.16	0.21	0.26	0.24
UMD beta	0.02	0.02	0.05	0.06	0.08	0.09
RMW beta	0.11	0.22	0.24	0.29	0.29	0.31
CMA beta	0.06	0.13	0.16	0.21	0.22	0.20

Three patterns are useful. First, the broad narrative distinctions are visible early. Operating-evidence attention rises from 0.01 with 10 dimensions to 0.23 with 40 dimensions and 0.43 with all 280 dimensions. Demand/revenue attention is even more front-loaded: the first 40 dimensions deliver an R^2 of 0.33, compared with 0.49 in the full embedding. The operating-evidence-minus-conditional-information contrast follows the same pattern, reaching 0.21 at 40 dimensions and 0.42 in the full embedding. These rows show that the embedding prefixes retain the paper's central distinction between concrete operating evidence and more conditional information.

Second, finer narrative distinctions require more of the embedding tail. Attention to conditional information is visible early but improves gradually, and guidance attention is weak throughout, reaching only 0.17 in the full embedding. The balance/capital versus commercial/product contrast is nearly absent at 10 dimensions but becomes strong at deeper truncations, rising to 0.20 at 40 dimensions and 0.43 in the full embedding. This ordering is the economically natural one: broad operating and demand information is front-loaded, while finer balance-sheet, financing, product, and customer distinctions require more coordinates.

Third, the external factor-beta rows do not all light up. Using monthly factor betas, the strongest exposure signal is profitability beta: RMW reaches an out-of-sample R^2 of 0.24 by 40 dimensions

and 0.31 in the full embedding. HML and CMA are also visible by 40 dimensions, with R^2 values of 0.16 and 0.16. Market beta and UMD beta are weaker, reaching only 0.15 and 0.09 in the full embedding. This selective pattern is useful negative evidence. The embeddings contain some persistent exposure structure, especially profitability, value, and investment exposure, but they do not mechanically encode every standard factor beta.

The dimensionality evidence therefore has a sharper interpretation than a generic "more dimensions are better" result. Table 9 shows an ordered representation: early prefixes preserve the main operating-information distinction and selected persistent exposures; later coordinates add finer narrative contrasts; momentum exposure remains weak.

5.6 Interpreting the Forecast-Value Cognitive Embedding

The previous subsection asks where predictive signal sits in the embedding coordinates. The interpretation in this subsection uses the object that actually forecasts the analyst's forecast. For the forecast-value model, that object is the horizon-specific cognitive embedding

$$e^V_{t,h} = \sum_j w^V_{jt,h} k_{jt}, \qquad \widehat{V}^{\text{std}}_{t,h} = g^V_h(e^V_{t,h}).$$

Thus the embedding is not a report-level average, a post-estimation PCA score, or a representation collapsed across forecast horizons. It is the same 280-dimensional attention-weighted call vector that is passed into the horizon-specific prediction head to recover the forecast value $V_{t,h}$. A report with several available horizons can therefore produce several forecast-value cognitive embeddings, one for each target horizon, because the attention weights $w^V_{jt,h}$ and prediction head g^V_h are horizon specific.

This target-consistent distinction is important. The forecast-error embedding answers a mistake question: which representation of the call helps explain why the forecast is too high or too low relative to realized earnings? The forecast-value embedding answers a level question: which representation of the call helps recover the analyst's stated forecast itself? If the goal is to interpret the embedding that "forecasts the forecast," the relevant object is $e^V_{t,h}$, not the forecast-error embedding $e^E_{t,h}$ and not an auxiliary report-level geometry probe.

The forecast-value embedding has three empirical properties. First, it contains substantial forecast-level information. Table 6 shows sparsemax analyst-disjoint R^2 values of 0.859 at 1y, 0.880 at 2y, 0.635 at 4y, 0.750 at 5y, and 0.652 for LTG, with the 3y horizon lower at 0.374. These are the R^2 values from the model map $g^V_h(e^V_{t,h})$, not from a separate interpretability regression. Second, the representation is not just a generic average of the call. Table 10 shows that replacing attention with a mean-pooled call embedding and passing it through the value-model head with estimated parameters held fixed is much weaker at every horizon. Panel A of Table 5 gives the

complementary query-shuffle test: holding the call and forecast value fixed, replacing the analyst's report query with another report query reduces R^2 by 0.277 at 1y, 0.313 at 2y, and 0.334 at 3y. The forecast-value embedding therefore depends on the analyst's written report as a retrieval cue, even though forecast levels also contain a large common firm and call component.

To interpret this embedding directly, I use the horizon-specific cognitive embeddings produced by the forecast-value attention model before the prediction head. The key question is whether the same representation $e^V_{t,h}$ that enters the value head is doing more than carrying a generic call average or a generic firm-year signal. Table 10 therefore collects the main identification checks for the value embedding in one place. *Head* R^2 is the performance of the nonlinear map $g^V_h(e^V_{t,h})$. *Frozen-random* R^2 replaces the learned sparsemax support with a same-k_{eff} random support and keeps the value head fixed. *Mean-head* R^2 passes the mean-pooled call embedding through the value head. *Query-shuffle drop* is the fall in R^2 when the report query is replaced by another report query while the call and target remain fixed. *Eff. keys* is $1/\sum_j (w^V_{jt,h})^2$, the effective number of call snippets in the sparsemax attention distribution.

Table 10: Evidence that the forecast-value cognitive embedding is predictive, report-conditioned, and sparse. $e^V_{t,h}$ is the cognitive embedding produced by the forecast-value attention model before the horizon-specific prediction head. *Head* R^2 is the analyst-disjoint out-of-sample R^2 from the value head $g^V_h(e^V_{t,h})$. *Frozen-random* R^2 uses the same effective number of call snippets but selects them randomly and keeps the value head fixed. *Mean-head* R^2 replaces learned attention with the mean call embedding and keeps the value head fixed. *Query-shuffle drop* is the loss in R^2 from replacing the report query with another report query. *Eff. keys* is $1/\sum_j (w^V_{jt,h})^2$.

Horizon	Head R^2	Frozen rand. R^2	Mean-head R^2	Query-shuffle drop	Eff. keys
YR 1	0.859	0.359	0.423	0.277	9.2
YR 2	0.880	0.365	0.424	0.313	8.9
YR 3	0.374	0.040	0.094	0.334	10.9
YR 4	0.635	0.317	0.356	0.234	8.0
YR 5	0.750	0.248	0.291	0.274	7.0
LTG	0.652	0.442	0.452	0.105	9.1

The table is more direct than a post-estimation PC probe. The forecast-value head is strong across horizons, while both nonsemantic references are materially lower. Frozen-random same-k_{eff} attention loses 0.21–0.52 R^2 points depending on horizon, and the mean-embedding fixed-head reference is much weaker than learned attention. The query-shuffle losses show that the report query matters even when the call and forecast value are held fixed. I therefore interpret $e^V_{t,h}$ as the target-consistent representation that recovers the analyst's stated forecast level.

Third, the forecast-value embedding is concentrated but not degenerate. Table 6 shows that sparsemax forecast-value attention uses roughly 6–10 effective call snippets depending on the hori-

zon, out of about 27–38 available snippets. This gives the embedding an economically readable structure: it is a weighted semantic summary of selected call content, not a one-snippet quote and not a flat document average. The same table shows that softmax often predicts forecast levels better, especially at 3y and 4y, but does so with much broader attention. I therefore use the sparsemax value embedding when the interpretation requires an explicit selected information set, while treating softmax as a useful predictive robustness check.

Extensive and intensive selection margins The extensive margin identifies how sparsity operates. For each taxonomy leaf c, define the forecast-value category attention share

$$A^V_{t,h,c} = \sum_{j \in \mathcal{J}_{t,c}} w^V_{jt,h},$$

where $\mathcal{J}_{t,c}$ is the set of candidate call snippets assigned to leaf c. I compute both the non-selection probability $\Pr(A^V_{t,h,c} = 0 \mid c \text{ present})$ and the conditional intensity $\mathbb{E}[A^V_{t,h,c} \mid A^V_{t,h,c} > 0,\ c \text{ present}]$. Aggregating these leaves into broad blocks is too coarse: many broad categories have similar non-selection rates because they mix screened-out background snippets with selected forecast-relevant snippets. Table 11 therefore reports a leaf-level selection frontier. Among leaves present in at least one percent of observation-horizons, Panel A shows the leaves most often assigned zero attention conditional on being available, while Panel B shows the leaves receiving the largest attention share conditional on being selected.

Table 11 and Figure 6 show that sparse forecast-value attention does not simply discard rare text, but the clearest pattern is leaf-specific and visible after grouping the frontier leaves into broader economic patterns. In the coarse bubble plot, firm and industry leaves are disproportionately in the upper-right region: conditional on appearing in the candidate call snippets, they are more likely to enter the sparsemax attention set and receive larger attention shares when selected. Macro and external leaves are more often in the lower-left region, meaning they are less often selected conditional on availability and tend to receive smaller conditional shares. Capital and markets leaves sit between these extremes. The screened-out frontier clusters around episodic or externally imposed states—exogenous shocks, event milestones, macro prices, policy and tax, operating frictions, contracting terms, and capital distribution. The high-intensity frontier instead clusters around persistent forecast primitives—demand and revenue, management outlook, product pipeline, operating economics, demand generation, customer loyalty, market/sector outlook, and related market-position information. This is a more modest but cleaner interpretation of sparsity. The value-attention model is not mechanically dropping all macro or financial information; rather, relative to macro/external background states, it puts more selection probability and more conditional attention mass on firm-specific operating mechanisms that map more directly into forecast

levels. The analogous forecast-error decomposition also separates non-selection from intensity. The main positive non-selection signals are leaf-specific, with the strongest interpretation coming from demand, cash/liquidity, regulation, investment, and product-market leaves rather than broad topic buckets. I next aggregate the same idea to the report level. Let $B_{t,h}$ denote the number of forecast-value topic leaves available in the candidate call snippets and $S_{t,h}$ the number receiving positive sparsemax attention. I define support compression as $C_{t,h} = 1 - S_{t,h}/B_{t,h}$, so higher values mean that a broad available menu is compressed into a narrower selected support.

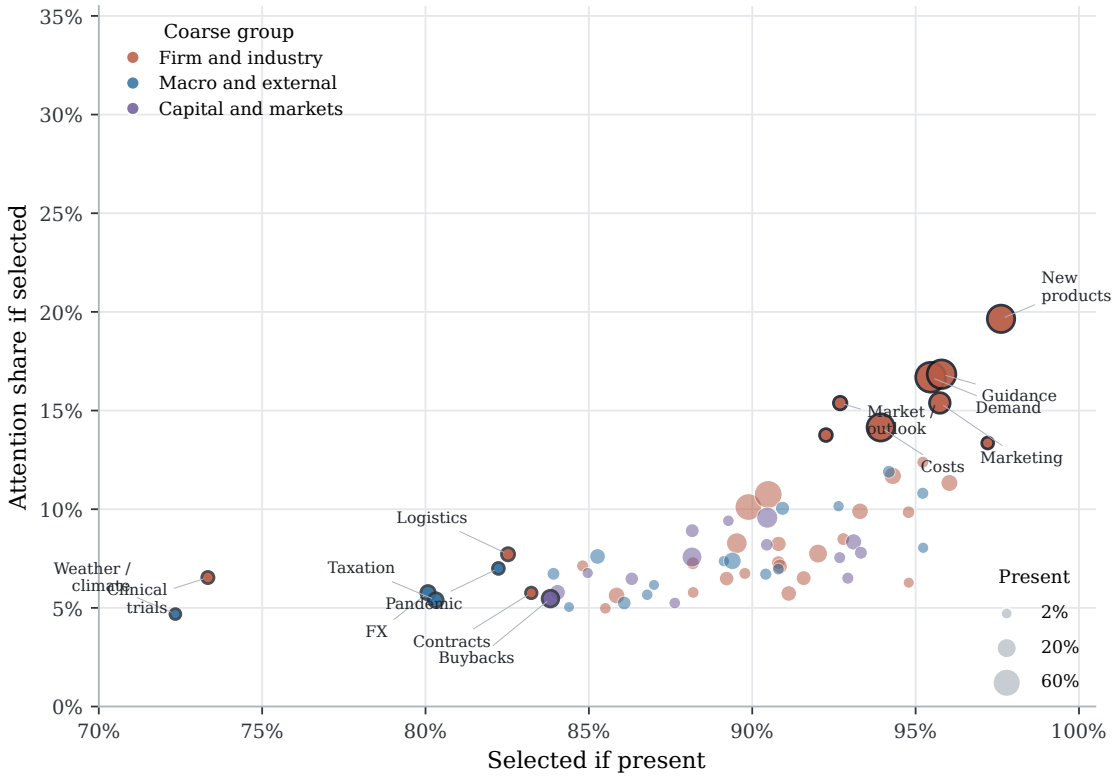
The key report-level intuition is that sparsity is not costly by itself. What matters is the combination of a broad available information menu and a narrow selected support.

Table 11: Forecast-value attention selection and intensity for the leaf-level frontier. The table uses the sparsemax forecast-value model. Panel A lists the taxonomy leaves most often screened out when they are present in the candidate call snippets. Panel B lists the leaves receiving the largest attention mass when they are selected. *Broad pattern* groups frontier leaves into coarser economic clusters. *Present* means that at least one candidate earnings-call snippet belongs to the leaf. *Ignored if present* is $\Pr(A^V_{t,h,c} = 0 \mid c \text{ present})$. *Share if selected* is $\mathbb{E}[A^V_{t,h,c} \mid A^V_{t,h,c} > 0,\ c \text{ present}]$. *Mean share* is the unconditional forecast-value attention share.

Taxonomy leaf	Broad pattern	Present	Ignored if present	Share if selected	Mean share
Environment: Weather and climate change	Exogenous shocks	1.1%	27.7%	4.7%	0.0%
Innovation / R&D: Clinical trials	Event milestones	2.4%	26.7%	6.5%	0.1%
Foreign exchange: Foreign exchange and currency rates	Macro prices	5.6%	19.9%	5.8%	0.3%
Domestic government: Taxation	Policy and tax	6.2%	19.7%	5.4%	0.3%
Crisis: Pandemic and public health issues	Exogenous shocks	2.1%	17.8%	7.0%	0.1%
Supply chain: Transportation and logistics	Operating frictions	3.6%	17.5%	7.7%	0.2%
Contracting: Contract terms and duration	Contracting	1.7%	16.8%	5.8%	0.1%
Payout: Stock buybacks and repurchases	Capital distribution	10.2%	16.2%	5.5%	0.5%
Product market competition: Customer loyalty and retention	Demand and market position	3.1%	7.7%	13.8%	0.4%
Industry: Market/sector outlook	Market outlook	4.2%	7.3%	15.4%	0.6%
Product market competition: Cost structure and efficiency	Operating economics	53.8%	6.1%	14.1%	7.2%
Customer: Customer demand/spending	Demand and revenue	71.2%	4.5%	16.7%	11.3%
Brand: Marketing and advertising	Demand generation	21.9%	4.3%	15.4%	3.2%
Governance: Management guidance	Management outlook	62.7%	4.2%	16.8%	10.1%
Brand: Brand equity and reputation	Other	2.0%	2.8%	13.4%	0.3%
Product: New products/services	Product pipeline	55.0%	2.4%	19.6%	10.6%

5.7 Cognitive Attention Counterfactuals

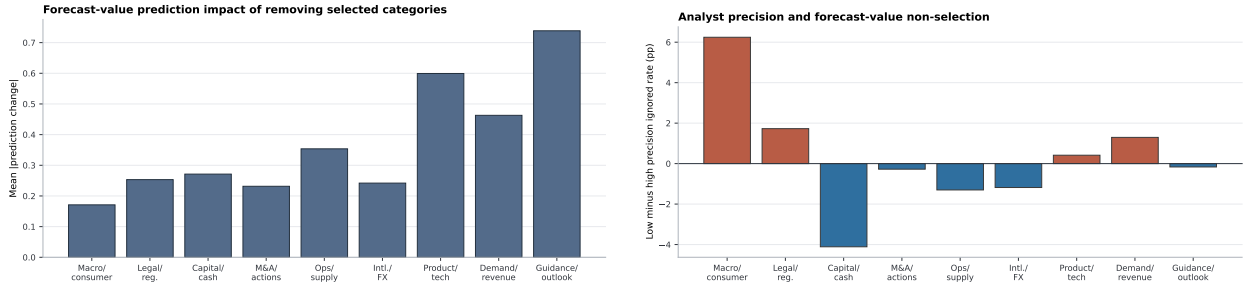
The selection frontier is economically useful only if selected leaves matter for the forecast representation. I therefore use cognitive attention counterfactuals: fixed-head exercises that ask how the forecast changes when the model is forced to remove text it had selected. Figure 7, Panel A, implements this logic. For each observation, horizon, and selected taxonomy leaf c, I remove that leaf's attention mass by setting the corresponding sparsemax weights to zero, renormalize the remaining selected call snippets, recompute the forecast-value embedding $e^V_{t,h}$, and pass the modified embedding through the already estimated value head g^V_h. The model is not retrained. The plotted statistic is the mean absolute change in the predicted forecast value induced by removing leaf c.



Leaves shown are present in at least 1% of observation-horizons. Bubble size is present share; dark outlines mark frontier leaves.

Figure 6: Bubble dot plot of forecast-value selection margins. Each point is a taxonomy leaf present in at least one percent of observation-horizons, excluding the residual other-operating fallback leaf. Fill color collapses leaves into three visual groups: firm and industry, macro and external, and capital and markets. The horizontal axis reports the extensive margin, $\Pr(A^V_{t,h,c} > 0 \mid c \text{ present})$, the probability that a present leaf receives positive sparsemax forecast-value attention. The vertical axis reports the intensive margin, $\mathbb{E}[A^V_{t,h,c} \mid A^V_{t,h,c} > 0,\ c \text{ present}]$, the mean attention share conditional on positive attention. Bubble size is proportional to the leaf's present share, and dark outlines mark the frontier leaves summarized in Table 11.

This exercise asks whether the selected leaves are merely visually salient in the attention map or whether they are prediction-moving components of the learned forecast representation.



(a) Prediction impact

(b) Precision and non-selection

Figure 7: Cognitive attention counterfactuals. Panel A reports the counterfactual prediction impact of selected forecast-value attention leaves: for each selected keyword-taxonomy leaf, the exercise removes that leaf's attention mass, renormalizes the remaining sparsemax weights, recomputes the forecast-value embedding, and reports the mean absolute change in the value-head prediction with estimated parameters held fixed. Panel B reports the low-minus-high precision difference in $\Pr(A^V_{t,h,c} = 0 \mid c \text{ present})$. Positive values mean low-precision analysts screen out the taxonomy leaf more often conditional on its presence; negative values mean high-precision analysts screen it out more often.

The largest removal effects line up with economically interpretable forecast primitives rather than with arbitrary text frequency. Industry life-cycle, management guidance, new products and services, customer demand, cost structure, market share, and cash/liquidity are among the leaves whose removal moves the forecast-value prediction most. Thus the attention map is not only sparse; the selected semantic leaves carry material forecast-level information after the prediction head is held fixed. Panel B uses the same extensive-margin object to preview the analyst-precision split: it asks whether low- and high-precision analysts differ in the leaves they screen out conditional on availability. I return to the economic interpretation of those precision gaps in Section 5.9.

The relationship between forecast-value embeddings and forecast-error embeddings is informative. The Supplemental Appendix transfer exercise in Section S.13 holds the value-trained attention map fixed, recomputes $e^V_{t,h}$ on the forecast-error sample, and passes the resulting embedding through the saved forecast-error prediction head without retraining. The transfer is meaningful, with mean analyst-OOS R^2 of 0.191, but it remains below the 0.393 achieved by the attention model trained directly on forecast errors. The implication is that $e^V_{t,h}$ captures much of the information analysts use to state forecasts, but forecast mistakes require additional target-specific attention structure. Put differently, forecast-value embeddings are the right object for interpreting the forecast itself; forecast-error embeddings are the right object for interpreting systematic over- or under-forecasting.

This also changes the role of narrative interpretation. Topic labels and named axes can still help

describe what the selected call snippets are about, but they should be applied after conditioning on the correct target. For forecast levels, the primitive is $e^V_{t,h}$: the attention-weighted call representation selected by the analyst's report and validated by its ability to recover $V_{t,h}$. I therefore use the forecast-value attention weights when comparing the selected information sets of high- and low-precision analysts, and return to the forecast-error attention weights only for exercises that condition directly on realized forecast errors.

5.8 Attention Concentration and Narrative Weights

The forecast-value embedding above interprets the representation that recovers forecast levels. I first use the same value-trained attention weights to compare the information sets selected by high- and low-precision analysts. I then return to the forecast-error attention weights for exercises that split observations by realized forecast errors. Those forecast-error exercises are most credible at 1y–3y, where the forecast-error validation evidence is strongest, but the same weights remain descriptively useful at the longer horizons as well.

Attention HHI is computed on normalized weights, $HHI = \sum_j w_j^2$.[3] In the forecast-error estimates, HHI ranges from about 0.11 to 0.21, with an average around 0.14, so the learned attention is meaningfully concentrated but far from degenerate. Appendix diagnostics report the corresponding forecast-error attention breadth on the analyst out-of-sample test set, including effective keys, nonzero attention share, top weight concentration, and HHI across horizons. Table 12 reports the same concentration diagnostics for the forecast-value model.

Table 12: Attention concentration statistics for the forecast-value multi-horizon attention model on the analyst out-of-sample test set. *Available call snippets* is the mean number of candidate call snippets. *Effective keys* $= 1/\sum_j w_j^2$ is the effective number of attended call snippets, and *Effective keys %* scales this by the number of available call snippets. *NZ frac.* is the fraction of call snippets with strictly positive attention weight, *Top wt.* is the weight on the most-attended call snippet, and *HHI* $= \sum_j w_j^2$ measures attention concentration.

Forecast	Avail. call snippets	Effective keys	Effective keys %	NZ frac.	Top wt.	HHI
YR 1	26.79	9.22	41.51	0.793	0.248	0.164
YR 2	26.64	8.93	40.98	0.808	0.259	0.170
YR 3	30.55	10.91	44.09	0.829	0.235	0.163
YR 4	34.25	7.98	32.39	0.719	0.316	0.225
YR 5	38.33	6.99	23.88	0.600	0.294	0.200
LTG	27.87	9.06	41.23	0.814	0.241	0.152

[3] Because attention weights sum to one, this is a unit-share HHI: flat attention over J call snippets gives $1/J$, a one-snippet spike gives 1, and multiplying by $10{,}000$ converts to the conventional antitrust scale.

5.9 Attention Shifts Across Analyst Precision

The next step asks whether the attention categories distinguish accurate and inaccurate forecasting behavior. The analyst-precision comparison uses the forecast-value attention model: analysts are classified by historical precision, with high-precision analysts in the lower tercile of pooled mean absolute scaled forecast error across LTG and 1y–5y forecasts and low-precision analysts in the upper tercile, but the plotted attention weights are the weights from the model trained to recover forecast values. I then separately return to the forecast-error attention model for observation-level low-error versus high-error tails, because that split conditions on the realized forecast error itself. The interpretation below places the most weight on the 1y–3y horizons, where the forecast-error validation evidence is strongest.

Main takeaway This subsection gives the cleanest behavioral evidence in the paper because the split is based on analysts' historical forecast precision, while the objects being compared are the model-implied information sets used to recover their stated forecasts. The clean distinction is not that good analysts attend to more information. They appear to select different kinds of information.

The memorable contrast is between *operating evidence* and *conditional signals*. Demand, revenue/orders, competition/pricing, margins, segment economics, product adoption, and customer usage are operating evidence because they are close to the earnings production function: quantities sold, realized pricing, customer traction, cost structure, and unit economics. By contrast, guidance/outlook, supply-chain disruptions, financing constraints, credit/funding, liquidity, and capital-allocation issues are conditional signals. They can be highly relevant, but their earnings implications depend on resolution, timing, financing conditions, and managerial interpretation. The argument is not that conditional signals are bad ex ante. The empirical result is that low-precision analysts select them more often, while high-precision analysts more often select direct operating evidence. Importantly, the result is mostly an extensive-margin difference: less accurate analysts are more likely to bring conditional signals into the attention set at all, rather than merely assigning slightly larger weights conditional on already attending to them.

Figure 8 shows the analyst-level version of this mechanism using the value-trained attention map and the same keyword category-label taxonomy as the benchmark sorts. The unconditional attention-share differences can look small because attention shares sum to one across many labels, so the figure decomposes the difference into the probability that a taxonomy leaf enters the attention set and the conditional share once it enters. The result is strongest on the entry margin. High-precision analysts are more likely to select new products/services, market share, customer demand/spending, cost structure and efficiency, and marketing and advertising. The largest selection gaps favoring high-precision analysts are customer demand/spending (8.5 percentage points), market share (8.3), new products/services (7.9), and marketing and advertising (6.8). These are

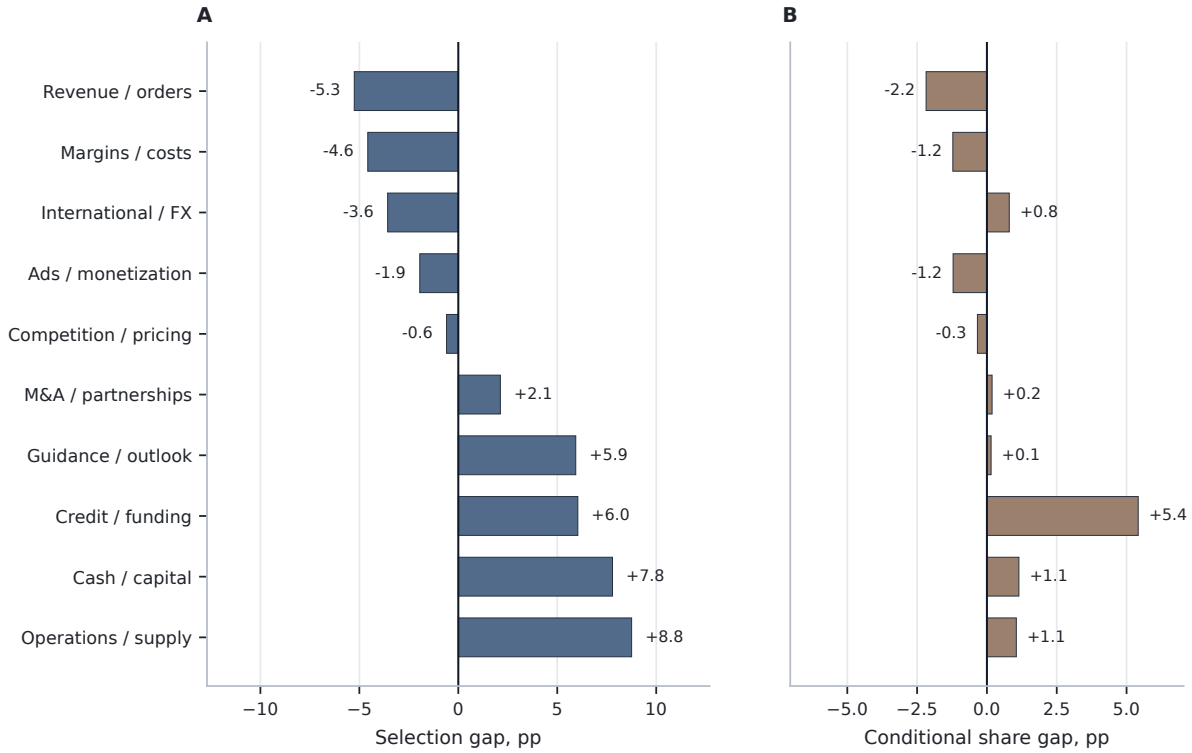


Figure 8: Forecast-value attention shifts between high- and low-precision analysts using the keyword category-label taxonomy. Bars report low-precision minus high-precision differences for the selected taxonomy leaves. Panel A decomposes the gap into the extensive margin, whether a category-label leaf enters the positive-attention set. Panel B reports the conditional share gap once the leaf is selected.

concrete product-market variables: evidence about whether customers are buying, whether the firm is gaining share, whether new products are taking hold, and whether costs convert into margins.

Low-precision analysts instead put more value-model attention on industry life-cycle discussion, management guidance, cash and liquidity, pricing strategy, and mortgage-rate material, with selection gaps of 6.7, 6.8, 6.0, 3.6, and 4.0 percentage points. This side of the figure is not noise. It is information whose cash-flow implication is more state-contingent: guidance must be interpreted, liquidity matters through constraints, and rate-sensitive demand depends on financing conditions. Thus the value model tells a sharper version of the paper's main behavioral story: high-precision analysts' forecast values are retrieved from direct operating and product-market evidence, while low-precision analysts' forecast values rely more on background conditions, guidance, and financing-condition labels. For asset-pricing applications, this distinction is useful because it separates attention to near-cash-flow fundamentals from attention to conditional state variables whose valuation implications require more modeling judgment.

Figure 7, Panel B, isolates the extensive margin even more directly. For each taxonomy leaf c, I condition on the leaf being present in the candidate call snippets and compute the probability that it receives zero sparsemax value-model attention, $\Pr(A^V_{t,h,c} = 0 \mid c \text{ present})$, separately for high- and low-precision analysts. The figure plots the low-minus-high difference in this ignored-if-present rate. Positive bars mean that low-precision analysts are more likely to screen out an available leaf; negative bars mean that high-precision analysts are more likely to screen it out. This is a stricter object than raw topic exposure because both groups are compared only when the information is available in the candidate call snippets.

The non-selection evidence reinforces the extensive-margin interpretation while also putting a boundary around it. Low-precision analysts are somewhat more likely to ignore available demand, brand, and taxation material, while high-precision analysts are more likely to screen out some capital-expenditure and cash/liquidity material. The result is therefore not a simple rule that better analysts always attend to every fundamental category. Instead, analyst precision is associated with a different filtering rule: high-precision analysts select direct operating evidence more reliably, while being more willing to discard some state-contingent or financing-related material when it is present.

Figure 9 moves from analyst types to realized forecast outcomes. It asks which taxonomy leaves would change predicted absolute forecast errors if their attention share were locally increased, holding the trained forecast-error head fixed. The same operating-versus-conditional distinction reappears, but now it is horizon specific. For one-year forecasts, negative cells indicate that high-error observations are especially sensitive to inventory management, cash and liquidity, hiring and layoffs, cost structure and efficiency, and supply-chain agility. This does not imply that

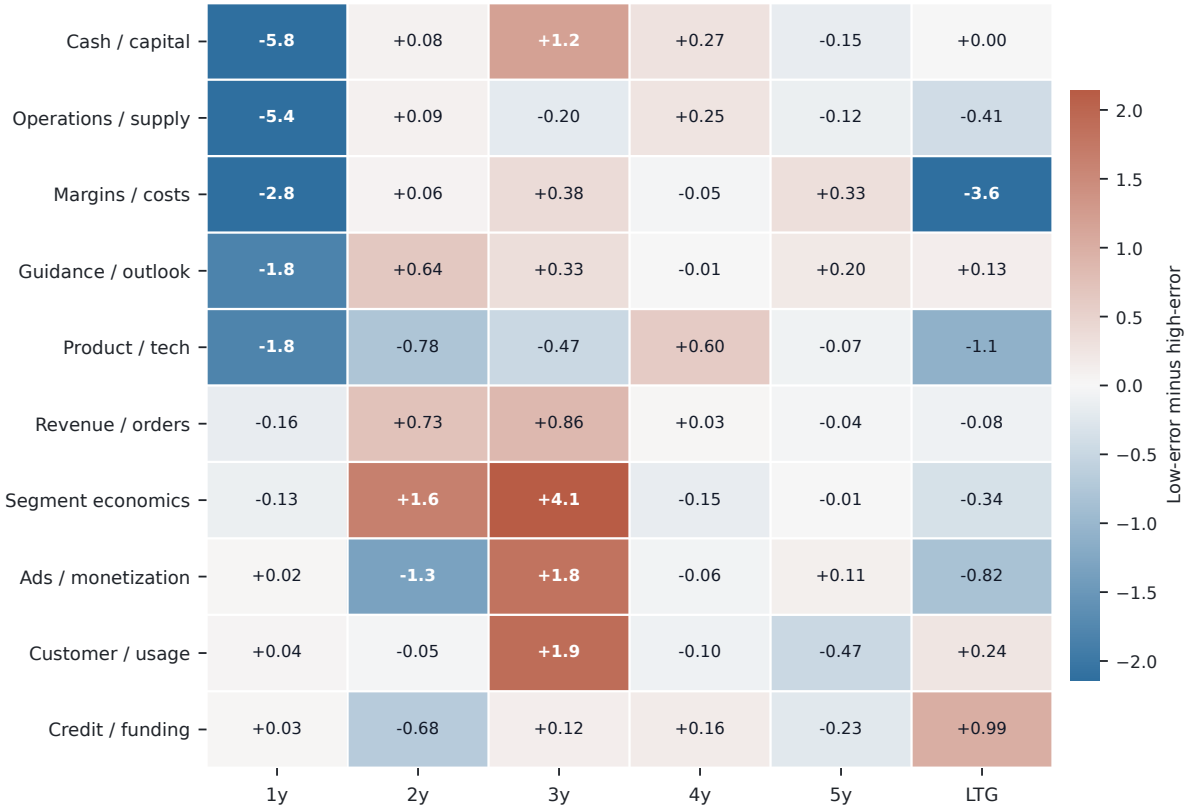


Figure 9: Attention-elasticity shifts between low-error and high-error forecasts using the keyword category-label taxonomy. For each observation t, horizon h, and taxonomy leaf k, I start from the forecast-error model's attention weights, increase the leaf's total attention share $S_{t,k,h} = \sum_{j \in \mathcal{J}_{t,k}} w_{jt,h}$ by 5 percent, shrink all non-k weights proportionally so attention still sums to one, recompute the attention-weighted embedding, and pass it through the forecast-error head with estimated parameters held fixed. The elasticity is $\varepsilon_{t,k,h} = \big(\Delta|\hat{y}_{t,h}|/|\hat{y}_{t,h}|\big)/\big(\Delta S_{t,k,h}/S_{t,k,h}\big)$, the percent change in predicted absolute forecast error per percent increase in leaf attention. Cells report the low-error-tail mean elasticity minus the high-error-tail mean elasticity within each horizon.

these labels are uninformative. Rather, in the realized high-error tail, forecast-error sensitivity is concentrated in labels whose earnings mapping is conditional on timing, execution, or financing constraints.

At the two-year horizon, high-error forecasts are more sensitive to new products/services and market share, while low-error forecasts are more sensitive to pricing strategy, management guidance, and product/service mix. At the three-year horizon, low-error forecasts load more on inventory management, cost structure and efficiency, customer loyalty and retention, and industry life-cycle information. The model's attention weights identify this as a horizon-matching problem: accuracy improves when attention shifts toward the evidence that is diagnostic for the forecast horizon, and the unified taxonomy makes the relevant labels explicit. Together, Figures 8 and 9 show that analyst skill is not just lower noise around the same information set. In the recovered attention map, high-precision forecasting is associated with a different selected-information set, tilted toward direct operating evidence and away from more conditional state variables.

Figure 10 asks whether the value-trained attention model changes the selected information set across forecast horizons. The category-label attention share is computed from the model's attention weights, not from raw topic frequency. For observation t, horizon h, and taxonomy leaf k,

$$S^V_{t,k,h} = \sum_{j \in \mathcal{J}_{t,k}} w^V_{jt,h},$$

where $w^V_{jt,h}$ is the attention weight assigned by the forecast-value model to call snippet j and $\mathcal{J}_{t,k}$ is the set of call snippets assigned to taxonomy leaf k. Thus, a leaf receives positive attention when at least one call snippet assigned to that leaf receives positive model attention. I compare short-horizon forecasts, $q = S$ for 1y–2y, with long-horizon forecasts, $q = L$ for 4y–5y and LTG. For each horizon group and taxonomy leaf, total attention can be written as

$$T_{q,k} = \mathbb{E}[S^V_{t,k,h} \mid q] = p_{q,k}\mu_{q,k},$$

where $p_{q,k} = \Pr(S^V_{t,k,h} > 0 \mid q)$ is the extensive margin and $\mu_{q,k} = \mathbb{E}[S^V_{t,k,h} \mid S^V_{t,k,h} > 0, q]$ is the intensive margin. The long-horizon minus short-horizon gap is decomposed symmetrically as

$$\Delta T_k = (p_{L,k} - p_{S,k})\bar{\mu}_k + (\mu_{L,k} - \mu_{S,k})\bar{p}_k,$$

where $\bar{\mu}_k = (\mu_{L,k}+\mu_{S,k})/2$ and $\bar{p}_k = (p_{L,k}+p_{S,k})/2$. The value-trained model has a different horizon signature from the forecast-error model. Moving from short to long horizons shifts attention toward industry life-cycle discussion, market share, M&A, CEO/senior-management discussion, marketing and advertising, product/service positioning, and product/service mix. Short horizons

receive relatively more attention to customer demand/spending, cost structure and efficiency, management guidance, and cash or liquidity. The decomposition shows that these value-model shifts are not just entry decisions: several of the largest long-horizon reallocations are driven by the conditional share once the label enters the attention set. Thus the value model appears to use attention to recover the level of the analyst's stated forecast from a horizon-specific information mix.

These gaps should be interpreted as reallocations within a fixed attention budget. Total model attention sums to one within an observation-horizon, so a positive value does not mean that long-horizon forecasts have more attention overall. It means that the value model allocates more of the fixed budget to that topic group when recovering long-horizon forecast levels.

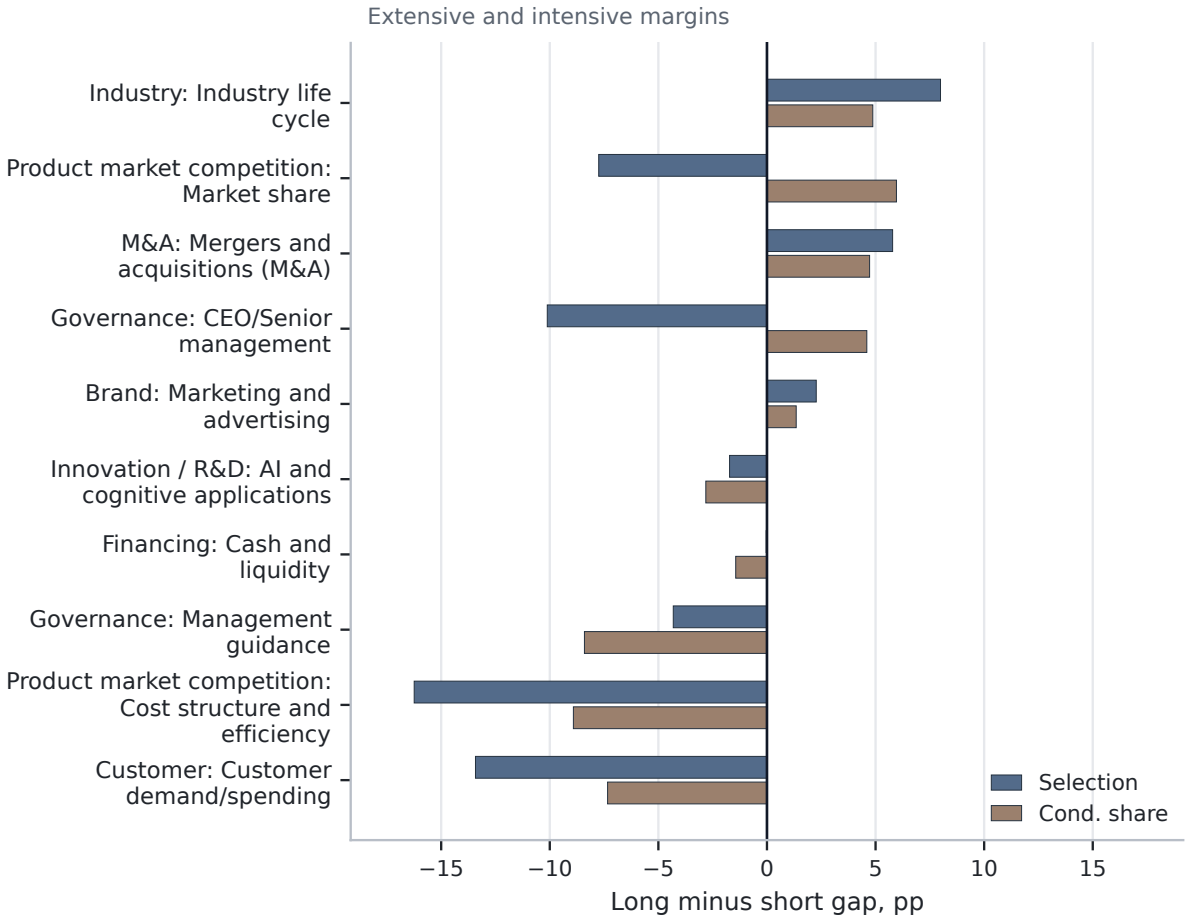


Figure 10: Forecast-value attention shifts across forecast horizons using the keyword category-label taxonomy. Bars compare long-horizon forecasts (4y, 5y, and LTG) with short-horizon forecasts (1y and 2y) for each selected taxonomy leaf. The dark component is the extensive-margin change: how much of the horizon difference comes from the leaf entering the positive-attention set more often. The light component is the intensive-margin change: how much comes from assigning more weight conditional on the leaf being selected. Positive values mean the leaf receives more forecast-value attention at long horizons; negative values mean it receives more attention at short horizons.

Figure 11 gives the forecast-error counterpart. It splits observations into low-error and high-error tails within horizon, then applies the same topic-share decomposition using attention weights from the forecast-error model. In that target-specific exercise, low-error forecasts place substantially more attention on the common-support topic universe at every horizon, with the largest gap at 2y. Almost all of the gap comes from the extensive margin. Thus, the main distinction in the forecast-error model is not that accurate forecasts finely tune weights within already-selected topics. Accurate forecasts are more likely to bring comparable earnings-relevant evidence into the attention set in the first place.

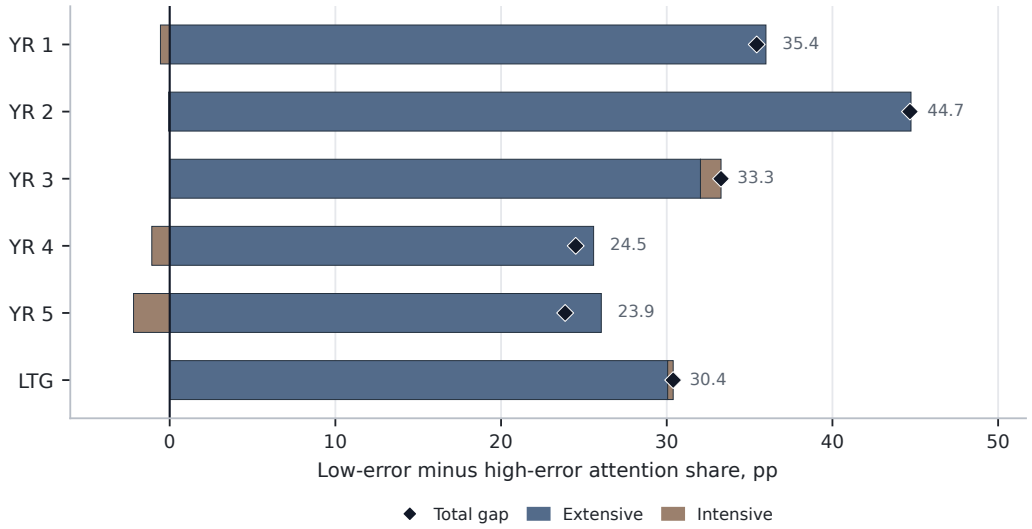


Figure 11: Attention gaps between low-error and high-error forecasts. For each forecast horizon, the figure compares the attention share assigned to topic labels that appear in both the low-error and high-error tails. The total bar is the low-error minus high-error attention gap. The dark component is the extensive margin, driven by whether a shared topic receives any positive sparsemax mass; the light component is the intensive margin, driven by the attention share conditional on selection. Positive values mean accurate forecasts allocate more attention to common-support earnings-relevant topics, mostly by selecting those topics in the first place.

Because the forecast-error exercise maps generated raw topic labels into the keyword taxonomy, the Appendix repeats the comparison without topic labels. That exercise clusters call-snippet embeddings directly and shows that the main concentration pattern is not purely an artifact of exact topic names.

5.10 Attention Temperature

The temperature parameter has a direct interpretation in the two constrained-attention benchmarks behind the paper. In Sims (2003), agents face a Shannon mutual-information cost. Solving the

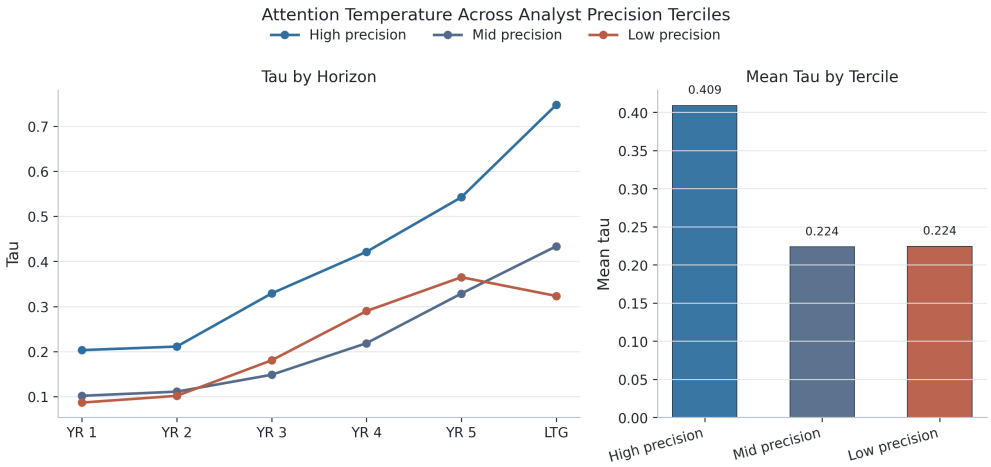


Figure 12: Estimated attention temperature across analyst-precision terciles for the forecast-value target. The exercise uses the full-separate sparsemax attention specification with raw forecast values as the target, while keeping the analyst-precision terciles defined from historical forecast errors.

constrained problem produces logit or softmax choice probabilities,

$$p(j) \propto \exp(\text{value}_j/\lambda),$$

where λ is the shadow cost of information: when information is cheap, the posterior is sharp; when information is costly, the posterior is diffuse. In Gabaix (2014, 2019), the regularization is instead sparse or quadratic, producing exact zeros in the attention allocation. The analogous parameter is the cognitive-cost parameter, often denoted κ. Thus the distinction is not that one parameter is behavioral and the other is not. Both are behavioral cost or discipline parameters; the difference is the penalty function. Entropic regularization yields softmax and dense attention, whereas sparse/quadratic regularization yields sparsemax-style allocations with zero weights. In the empirical model, τ_h plays the corresponding role: it governs how strongly semantic score differences are converted into an attention allocation. I interpret it as a reduced-form attention-cost or retrieval-sharpness parameter, not as a separately identified structural welfare primitive.

Figure 12 focuses on the identified temperature parameter τ_h from the forecast-value target. The horizon pattern is the cleanest Sims-style comparative static. Lower τ_h means that semantic distinctions are converted more aggressively into attention weights. The short horizons, especially *1y* and *2y*, have low estimated τ_h, so their forecast-value representations are sharp: prediction is driven by a relatively small set of diagnostic disclosures such as near-term guidance, margin commentary, orders, contracts, or demand updates. In this case the information environment itself rewards concentration. By contrast, τ_h rises at longer horizons. Three- to five-year forecasts and LTG forecasts are less likely to be pinned down by a single call snippet; they depend on many slowly evolving

fundamentals. The model therefore recovers softer attention because the long-horizon information environment rewards diffusion across many moderately informative snippets.

The analyst-precision pattern should be read more cautiously. Higher-precision analysts have larger τ_h values on average, about 0.409, compared with about 0.224 for mid- and low-precision analysts. This does *not* mean that low-precision analysts pay less attention in total, nor does it imply that high-precision analysts face higher structural information costs. Attention weights always sum to one, and τ_h is estimated inside a predictive retrieval model. The result says that low-precision analysts behave *as if* their report-query representation makes sharper distinctions among call snippets, producing a narrower selected-information set. High-precision analysts, by contrast, behave *as if* they use smoother attention and integrate a broader set of semantically relevant disclosures. A useful interpretation is epistemic robustness rather than welfare: better forecasting is associated with less brittle evidence integration, while lower precision is associated with a more selective and potentially overconfident attention profile.

5.11 Benchmark Cross-Sectional Sorts

Before using forecast-value attention to interpret factor narratives, I first ask a more mechanical benchmark question: do analyst mental-model measures vary systematically across standard characteristic portfolios? For each report i, I summarize the value-forecast transformer by three families of variables. First, net sentiment is

$$S_i = \frac{1}{|\mathcal{H}_i|} \sum_{h \in \mathcal{H}_i} \left(A_{i,h}^{V,+} - A_{i,h}^{V,-} \right),$$

where $A_{i,h}^{V,+}$ and $A_{i,h}^{V,-}$ are the positive- and negative-tone attention shares at horizon h, aggregated across attended topics. Second, I measure attention breadth using normalized entropy and sparsity of the forecast-value attention distribution. Third, following Bastianello, Décaire, and Günzel (2024), I allocate every raw topic, and hence every attention weight attached to that topic, to exactly one leaf in a pre-specified keyword taxonomy. Let $\ell(j)$ denote the raw topic label attached to snippet j, and let $d(\ell) = (g, m) \in \mathcal{G} \times \mathcal{M}_g$ be the deterministic assignment for raw topic ℓ to parent category g and leaf label m. The parent taxonomy set $\mathcal{G}$ contains pricing, product-market competition, customer, brand, product, supply-chain, innovation/R&D, corporate investment, production/utilization, financing, payout, M&A, contracting, governance, labor market, environment, legal, regulation, domestic government, housing market, demographic, crisis, raw input, GDP growth, inflation, foreign exchange, international, industry, discount-rate, systematic-risk, interest-rate, and valuation categories. Each parent category g contains economically named leaf labels $\mathcal{M}_g$, such as product/service positioning and differentiation, customer demand/spending, or man-

agement guidance. The category-label forecast-value attention share is

$$A^V_{i,h,g,m} = \sum_{j:\, d(\ell(j))=(g,m)} w^V_{ijh}, \qquad g \in \mathcal{G},\ m \in \mathcal{M}_g.$$

The assignment rule is intentionally mechanical and keyword based: raw-topic keyword matches are used first; if the raw topic has no taxonomy keyword, the coarser economic-core label supplies a fixed fallback leaf. This makes the category-label rows mutually exclusive and exhaustive for the attention-weight rows.

Figure 13 sorts observations into monthly q20/q80 tails on CAPM beta, market-to-book, size, momentum, realized next-year return, investment, and profitability. Each cell reports the high-minus-low difference in the mental-model measure after absorbing SIC2 industry-year means. The exercise is descriptive. It does not ask whether the measures predict returns; it asks whether analysts' value-forecast representations differ in recognizable ways across standard firm types. The main heatmap reports all taxonomy leaves with at least one statistically nonzero sort difference of at least two percentage points, ordered by the largest such difference for readability. Appendix Section A.6 reports the complete category-label taxonomy.

Table 13: Selected benchmark sort differences using signal-bearing taxonomy category-label rows

Measure	CAPM beta	Market-to-book	Size	Momentum	Next-year return	Investment	Profitability
Governance: Management guidance	1.42	-0.70	-11.99*	-0.73	0.66	-0.39	-4.93*
Product: New products/services	-7.12*	5.24*	11.93*	-1.95*	-1.81*	-0.89	2.64*
Customer: Customer demand/spending	-1.33*	2.39*	5.32*	1.85*	-0.59	-1.11	3.85*
Industry: Industry life cycle	4.47*	-2.00	-5.19*	3.33*	2.76*	4.99*	-4.92*
Innovation / R&D: Innovation and R&D	0.64	-2.25*	-4.09*	-1.68*	0.85	1.40*	-2.30*
Product market competition: Market share	-1.48*	-1.67*	3.91*	-2.11*	0.12	0.21	1.89*
Brand: Marketing and advertising	0.79*	-3.54*	1.39*	-1.24*	-2.45*	1.20*	2.56*
Product market competition: Cost structure and efficiency	-2.68*	0.57	3.33*	-0.17	3.47*	-1.93*	-3.21*
Financing: Cash and liquidity	0.30	-1.93*	-3.38*	-1.26*	-0.34	-0.33	-1.51*
Governance: CEO/Senior management	-2.85*	0.15	2.25*	-2.39*	-0.57*	-2.82*	0.39
Supply chain: Inventory management	0.51*	-0.70*	-2.14*	-0.38*	0.17	0.16	-0.18

Notes: Entries are high-minus-low q20/q80 sort differences in percentage points after residualizing by SIC2 industry-year. Category-label rows use the exclusive keyword-based raw-topic taxonomy assignment described in the text, are displayed as parent category followed by leaf label, and are included when at least one sort difference is statistically nonzero and at least two percentage points in absolute value. Asterisks mark 95% intervals excluding zero. The next-year-return sort is descriptive and uses realized twelve-month returns after the report.

Several benchmark patterns are economically intuitive. Large firms receive much less net positive sentiment and less attention to management guidance and cash or liquidity than small firms, but more attention to customer demand, new products and services, market share, and cost structure. High market-to-book firms receive more attention to new products and services, but less attention to marketing, innovation/R&D, cash or liquidity, and market share. High beta firms receive less attention to new products and services and more attention to industry life-cycle information. Profitable firms receive more attention to customer demand and marketing, but less attention to management guidance, industry life-cycle information, cost structure, cash or liquidity, and innovation/R&D. The descriptive point is not that these attention measures are new priced factors.

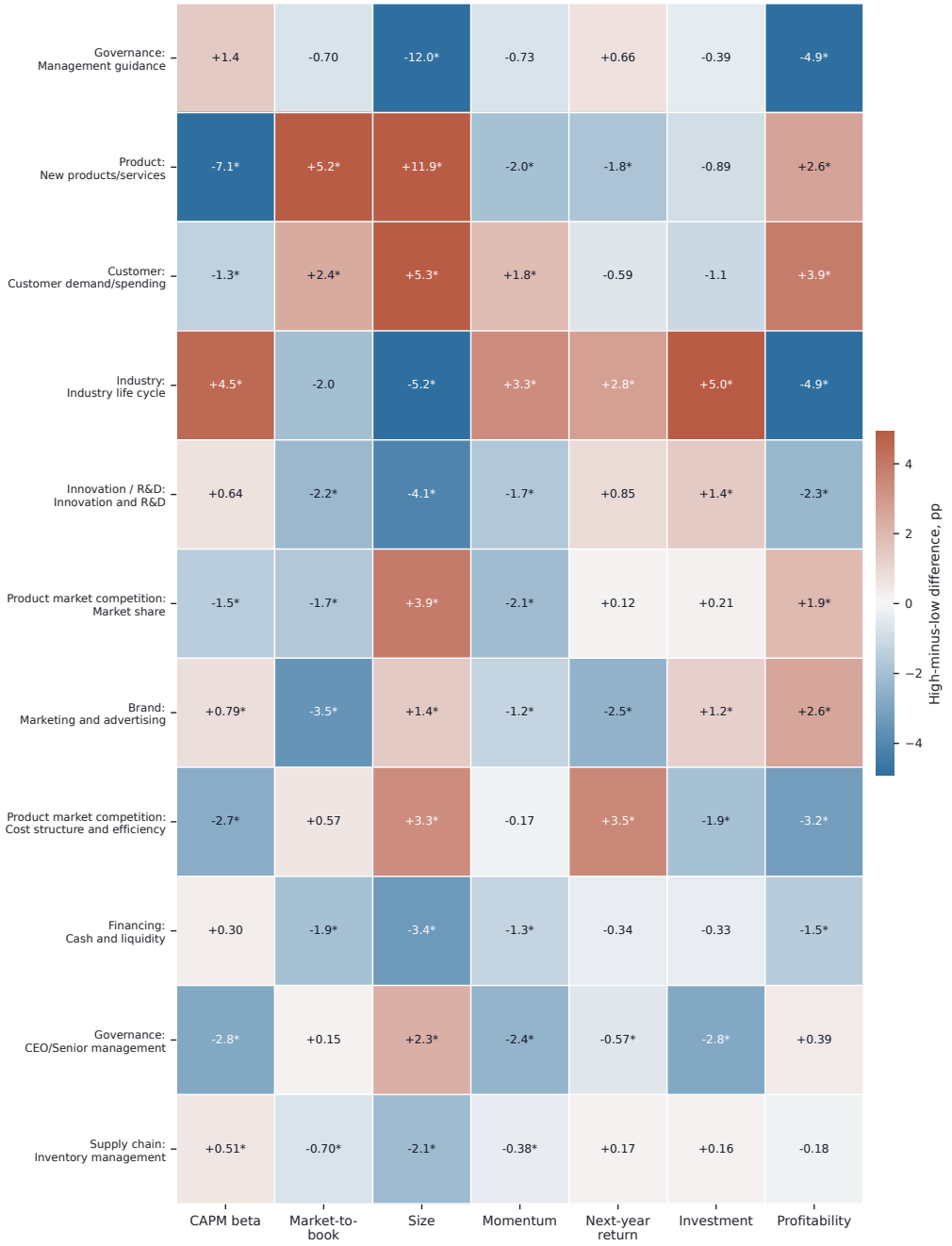


Figure 13: Benchmark cross-sectional sorts of forecast-value mental-model measures using mutually exclusive taxonomy category-label rows. Columns sort reports into monthly q20/q80 high and low tails on standard characteristics; entries are high-minus-low differences in percentage points after absorbing SIC2 industry-year means. Rows keep all taxonomy leaves with at least one statistically nonzero sort difference of at least two percentage points. Each raw topic, and therefore each attention weight attached to that topic, is assigned to exactly one keyword-based taxonomy leaf, displayed as parent category followed by leaf label. CAPM beta is a trailing 36-month market beta. Next-year return is the realized twelve-month return after the report and is included as an ex-post descriptive sort. Asterisks mark 95% intervals excluding zero.

It is that the value-transformer mental model varies systematically across familiar cross-sectional portfolios before any return-pricing interpretation is imposed.

6 Conclusion

Analysts observe the same earnings calls and produce different forecasts. This paper measures that divergence by using each analyst's written report as a behavioral query over the matched call. A query/key attention model recovers which call snippets the report most closely tracks, and the resulting attention-weighted representation explains the term structure of EPS forecasts, long-term growth forecasts, and forecast errors out of sample on analysts not seen in training.

Three findings discipline the interpretation. Learned attention improves on sparsity-matched random pooling and recovers an interpretable map from attended disclosures to forecast values and forecast errors. The query-shuffle falsification shows that matched analyst text carries real information: replacing an analyst's report with another report from the test set lowers performance substantially at the one, two, and three year horizons, so the model is not extracting generic call-level predictability. Forecast values and forecast errors are related but distinct. Holding the value-trained attention map fixed and passing the resulting embedding through the saved forecast-error head recovers part, but not all, of the forecast-error signal, so forecasts and forecast mistakes are not summarized by the same selected information set.

The patterns are informative. Short-horizon forecasts are recovered from a concentrated set of operating disclosures, while long-horizon forecasts require broader semantic aggregation. High-precision analysts' forecasts align with realized operating evidence such as demand, market share, customer behavior, and product traction, while less precise analysts' forecasts align more with conditional information whose earnings implications depend on resolution, timing, and managerial interpretation. The difference is primarily on the extensive margin. Accuracy comes from bringing the right kind of evidence into the selected set, not from fine reweighting once it is in.

The framework is portable. Any setting with a common information environment, behavioral text written by an agent, and a downstream outcome admits the same inversion design. The behavioral extensions in Appendix A.2, covering episodic memory, diagnostic representativeness, salience, prior anchoring, sticky updating, and confirmation bias, show how the same architecture nests canonical theories of expectation formation, each identified through a different empirical moment. The within-call estimates here are the first step in that program.

A Appendix

A.1 Snippet Extraction Prompts

This appendix reports the prompts used to convert earnings-call transcripts and analyst reports into the snippet inputs used throughout the paper. The prompts are designed for extraction and segmentation rather than interpretation. They instruct the model to preserve the source speaker's language, tone, and slant; to remove only non-substantive boilerplate; and not to invent content or merge distinct themes. The JSON topic labels are used for descriptive diagnostics, but the baseline attention model uses the snippet embeddings rather than the topic labels.

Call system prompt

```
You are a financial analyst extracting key management insights and business outlook from
    earnings call transcripts. Focus on extracting management's specific views, guidance, and
    strategic insights about the company's performance and future prospects.
```

Call user prompt

```
Task:
1. Identify distinct topics/themes in the call (e.g., "Revenue Growth", "Margin Pressure", "
    Market Competition", "Management Strategy", "Product Development", etc.)
2. For each topic, extract all key management insights that reflect their views and guidance
3. Keep each snippet concise but provide enough context (2-4 sentences, or one longer
    paragraph if needed for clarity)
4. Keep management's language and tone (optimistic/cautious/uncertain). Do NOT neutralize tone
    . Do NOT paraphrase into generic finance-speak
5. Remove only the following:
  - Opening/closing formalities ("Thank you for joining us", "That concludes our call")
  - Operator instructions and procedural language
  - Pure Q&A transitions ("Next question", "Thank you for the question")
  - Pure filler with no business content ("We're excited about the future", "We remain focused
      on execution" -- drop it unless it names specific drivers or metrics)
6. Do NOT invent content. Do NOT merge multiple themes into one snippet

Output format:
Return a JSON object with topics as keys and arrays of snippets as values:
{
  "Topic 1 Name": ["snippet 1", "snippet 2"],
  "Topic 2 Name": ["snippet 1", "snippet 2", "snippet 3"],
  "Topic 3 Name": ["snippet 1"]
}

Do not include any commentary or explanation outside the JSON.

Example of what to KEEP:
- "We expect gross margin to be about 36%, down from 41.7% in the March quarter and reflecting
     approximately $36 million related to stock-based compensation expense."
- "The demand in the U.S. was much, much stronger than we predicted. And so regrettably, we
    had to push out the international launch to be able to launch the 3G unit in the U.S."
```

```
- "We think the market for the iPad will be large, and we want to capitalize on our first
    mover advantage. We have been very aggressive with pricing and are delivering tremendous
    value to customers."

Example of what to DROP:
- "Thank you for joining us today."
- "Next question please."
- "We remain focused on execution and delivering value to shareholders."

Now here is the call transcript:

{text}
```

Report system prompt

```
You are a financial analyst extracting key analyst beliefs and interpretations from corporate
    communication. Focus on extracting the analyst's specific views, interpretations, and
    predictions about the company's situation.
```

Report user prompt

```
Task:
1. Identify distinct topics/themes in the text (e.g., "Revenue Growth", "Margin Pressure", "
    Market Competition", "Management Strategy", etc.)
2. For each topic, extract all key belief snippets that reflect the analyst's interpretation
3. Keep each snippet concise but provide enough context (2-4 sentences, or one longer
    paragraph if needed for clarity)
4. Keep the analyst's language and slant (bullish/bearish/uncertain). Do NOT neutralize tone.
    Do NOT paraphrase into generic finance-speak
5. Remove only the following:
  - Compliance / disclaimer language
  - Rating boilerplate ("Overweight, PT $142" etc.)
  - Ticker insert garbage
  - Pure filler with no economic content ("we remain constructive on management," "we like the
      story near-term" -- drop it unless it names an actual driver)
6. Do NOT invent content. Do NOT merge multiple themes into one snippet

Output format:
Return a JSON object with topics as keys and arrays of snippets as values:
{
 "Topic 1 Name": ["snippet 1", "snippet 2"],
 "Topic 2 Name": ["snippet 1", "snippet 2", "snippet 3"],
 "Topic 3 Name": ["snippet 1"]
}

Do not include any commentary or explanation outside the JSON.

Example of what to KEEP:
- "We think margin pressure will persist into 2H because input cost inflation is not being
    fully passed through. This reflects both competitive pricing pressures and the company's
    inability to fully offset rising costs through price increases."
```

```
- "Management backed away from the volume guide, which we see as an admission that demand is
    weakening in Europe. The revised guidance suggests underlying demand trends are worse than
     initially expected, particularly in key European markets."
- "We expect pricing power to fade next quarter as competitive intensity increases and
    customer price sensitivity rises. This could pressure margins despite management's efforts
     to maintain pricing discipline."

Example of what to DROP:
- "We reiterate OW and $142 PT."
- "We remain constructive on the name."
- "See important disclosures at the end of this note."

Now here is the report text:

{text}
```

A.2 Behavioral extensions of the attention rule

The baseline attention rule treats the within-call logit $z_{ijt,h} = \tilde{q}_{it,h}^{\top}\tilde{k}_{jt,h}/\tau_h$ as a pure semantic match scaled by the temperature τ_h. This subsection describes six extensions that map directly onto canonical theories of expectation formation. Each extension modifies either the logit $z_{ijt,h}$, the key set, the pooling rule that produces the cognitive embedding $e_{t,h}$, or the training loss. Each nests the baseline at a specific parameter value or data restriction, so the relevant cognitive distortion is identified as a deviation from that null.

A.2.1 Memory as episodic retrieval

The baseline specification restricts the key set K_t to call snippets from the matched earnings call. This makes the recovered cognitive embedding a within-call selection rule. A natural extension widens the key set to include past episodes: earlier calls for the same firm, prior reports written by the same analyst, peer-firm calls, and contemporaneous news. With this extension, the attention layer no longer summarizes a single document. It becomes a retrieval mechanism over an episodic memory store, scored by how well each past snippet matches the analyst's current concern.

For analyst–firm observation t, define the augmented key set

$$K_t^{\text{aug}} = K_t \cup \bigcup_{s \in \mathcal{M}(t)} K_s,$$

where $\mathcal{M}(t)$ indexes past episodes admissible when the report is written and K_s is the snippet collection for past episode s. Each past snippet $j \in K_s$ carries an episode timestamp $\tau(j)$. The relevant time variable is the age of the snippet's source episode relative to the report,

$$\Delta_{jt}^{\text{age}} = \frac{\text{date}(t) - \tau(j)}{T},$$

where T is a normalization horizon. The augmented logit is

$$z_{ijt,h} = \frac{\tilde{q}_{it,h}^{\top}\tilde{k}_{jt,h}}{\tau_h} - \delta_h \left(\Delta_{jt}^{\text{age}}\right)^{\alpha_h}, \qquad \delta_h \geq 0, \quad \alpha_h \in (0, 1].$$

Two forgetting parameters now have separate roles. δ_h controls the scale of episode-age discounting, while α_h controls the curvature of the forgetting curve. When the key set contains only one source episode, such as the matched call in the baseline, Δ_{jt}^{age} is constant across snippets and the age penalty drops out of the allocation rule. The form also nests a linear logit penalty when $\alpha_h = 1$ and a sublinear, power-law-style forgetting curve when $\alpha_h < 1$, consistent with evidence on slow memory decay (Anderson and Schooler, 1991; Wachter and Kahana, 2024). δ_h is identified from variation in episode age within K_t^{aug}, while α_h is identified from curvature in the attention drop-off across episode-age bins.

This formulation links the paper's attention model to episodic retrieval in cognitive psychology. In temporal-context models, items are encoded with a slowly drifting context, and recall is a similarity-weighted competition between cue match and context overlap (Howard and Kahana, 2002; Polyn, Norman, and Kahana, 2009). The augmented logit has the same structure: the semantic channel $\tilde{q}^{\top}\tilde{k}$ asks whether a past snippet matches the current report cue, while $-\delta_h(\Delta^{\text{age}})^{\alpha_h}$ penalizes context drift. Older snippets can still be retrieved, but they must be more semantically aligned to compete.

The extension also connects to finance models in which beliefs are shaped by selective recall rather than full-information processing. Wachter and Kahana (2024) develop a retrieved-context theory of financial decisions, Nagel and Xu (2022) document fading-memory dynamics in expectations and asset prices, and Malmendier and Nagel (2011, 2016) show that experienced macroeconomic histories shape later financial decisions. In this interpretation, the present paper estimates the within-call version of a broader memory model: τ_h governs retrieval sharpness, while the proposed pair (δ_h, α_h) would govern the scale and curvature of fading across episodes.

The same idea appears in machine learning. Retrieval-augmented language models improve prediction by searching an external evidence store rather than relying only on the current context window (Lewis, Perez, Piktus, Petroni, Karpukhin, Goyal, Küttler, Lewis, Yih, Rocktäschel, Riedel, and Kiela, 2020; Borgeaud, Mensch, Hoffmann, Cai, Rutherford, Millican, van den Driessche, Lespiau, Damoc, Clark, et al., 2022), and modern Hopfield-network results show that transformer attention can be interpreted as continuous-state associative memory retrieval (Ramsauer, Schäfl, Lehner, Seidl, Widrich, Gruber, Holzleitner, Adler, Kreil, Kopp, Klambauer, Brandstetter, and Hochreiter, 2021). The economic point is that attention and memory are complements. When diagnostic evidence lives inside the matched call, the baseline within-call attention model is the right inversion target. When diagnostic evidence lives in past episodes, the augmented memory

model is the right target.

A.2.2 Diagnostic expectations

Kahneman and Tversky (1972) introduced the representativeness heuristic: agents judge how typical an outcome is of a class by how much it stands out relative to a baseline. Bordalo, Gennaioli, and Shleifer (2018) formalize this logic as diagnostic expectations, in which agents overweight states whose conditional likelihood has risen relative to a reference distribution. Bordalo, Gennaioli, Ma, and Shleifer (2020) document that professional macroeconomic forecasters overreact to news in this direction, and Bordalo, Gennaioli, La Porta, and Shleifer (2024) extend the framework to financial-market expectations.

The diagnostic extension amends the attention model by adding a representativeness term to the logit:

$$z_{ijt,h} = \underbrace{\frac{\tilde{q}_{it,h}^{\top}\tilde{k}_{jt,h}}{\tau_h}}_{\text{semantic match}} + \theta_h \log \frac{p\left(\tilde{k}_{jt,h} \mid \mathcal{R}_t\right)}{p\left(\tilde{k}_{jt,h} \mid \mathcal{B}_t\right)}, \tag{1}$$

where $\mathcal{R}_t$ is a recent reference window, such as the empirical distribution of call-snippet embeddings in the trailing 30–90 days, and $\mathcal{B}_t$ is a longer baseline window. The density ratio is positive for snippets whose semantic direction has become more frequent recently than historically and negative for snippets whose direction has become relatively rarer. The parameter θ_h measures the strength of diagnostic distortion. When $\theta_h = 0$, the logit collapses to the baseline. When $\theta_h > 0$, the cognitive embedding is pulled toward recently representative embedding directions after holding fixed the rational semantic match $\tilde{q}_{it,h}^{\top}\tilde{k}_{jt,h}$.

A.2.3 Bayesian updating with a prior anchor

The baseline pooling rule $e_{t,h} = \sum_j w_{jt,h} k_{jt}$ treats the cognitive embedding as a function of the current call alone. This implicitly imposes an uninformative prior. A Bayesian forecaster would combine new information with a prior, and the strength of that prior governs how sluggishly beliefs respond to new disclosures. Coibion and Gorodnichenko (2015) develop the canonical information-rigidity test by asking whether forecast errors are predictable from past forecast revisions. Bouchaud, Krueger, Landier, and Thesmar (2019) document that equity analysts' return expectations are sticky in this sense. Nagel and Xu (2022) model belief formation as a fading-memory average of past data, which is observationally close to Bayesian updating with a prior estimated from history.

This extension amends the pooling stage rather than the logit. Let $\mu_{a,f}^{\text{prior}}$ denote a prior embedding for analyst a on firm f, constructed as the average of her cognitive embeddings on f over the

previous K calls. The Bayesian-pooling cognitive embedding is

$$e_{t,h} = \lambda_{a,f,h}\mu_{a,f}^{\text{prior}} + (1 - \lambda_{a,f,h}) \sum_j w_{jt,h} k_{jt}, \qquad \lambda_{a,f,h} \in [0, 1]. \tag{2}$$

The prior weight $\lambda_{a,f,h}$ measures how sluggishly the analyst updates from the matched call. When $\lambda_{a,f,h} = 0$, the embedding is built entirely from the current call, recovering the baseline. When $\lambda_{a,f,h}$ approaches one, the analyst effectively restates her prior view. Identification comes from within-analyst-firm persistence of $e_{t,h}$ across consecutive calls. A richer version would let $\lambda_{a,f,h}$ depend on the informativeness or diffuseness of the call; an empirical prior weight that fails to respond to call informativeness is itself a behavioral deviation.

A.2.4 Sticky inattention

Sims (2003) introduces rational inattention, in which information processing is costly and agents optimally smooth information acquisition. Maćkowiak and Wiederholt (2009) use rational inattention to microfound sticky prices: firms re-optimize infrequently because attention is scarce. Applied to belief formation, the implication is that agents revise their cognitive state only when the marginal information gain exceeds a cognitive cost, producing infrequent but larger revisions.

The sticky-inattention extension amends the training objective. The forward pass can remain unchanged, but the loss penalizes large changes in the cognitive embedding from one call to the next:

$$\mathcal{L}_t^{\text{sticky}} = \mathcal{L}_t^{\text{pred}} + \kappa_h \left\| e_{t,h} - e_{t-1,h} \right\|^2, \tag{3}$$

where $e_{t-1,h}$ is the cognitive embedding for the same analyst–firm one call earlier and $\mathcal{L}_t^{\text{pred}}$ is the masked Smooth-L1 loss from the main specification. The parameter κ_h is the per-unit cost of belief revision, with $\kappa_h = 0$ recovering the baseline. Identification requires a panel with multiple calls per analyst–firm and exploits the autocorrelation of $e_{t,h}$. A large κ_h produces lumpy, staircase-shaped belief paths; a small κ_h lets beliefs fluctuate more freely with each new call.

A.2.5 Salience

Bordalo, Gennaioli, and Shleifer (2012) develop salience theory, in which attention is drawn to options or attributes that contrast most sharply with a reference point. Salience is distinct from diagnosticity: diagnosticity uses a temporal reference, recent versus usual, whereas salience uses a cross-sectional or context-specific reference. Bordalo, Gennaioli, and Shleifer (2013) embed salience in an asset-pricing framework and show that contrast-based attention generates predictable distortions in expected returns.

In the present setting, the natural reference for snippet j is firm f's historical baseline embedding, $\bar{k}_f$, defined as the average call-snippet embedding across that firm's prior earnings calls. The

salience extension amends the logit with a contrast term:

$$z_{ijt,h} = \frac{\tilde{q}_{it,h}^{\top}\tilde{k}_{jt,h}}{\tau_h} + \sigma_h \left\| \tilde{k}_{jt,h} - \bar{k}_f \right\|. \tag{4}$$

A snippet close to the firm's typical content receives little or no salience boost. A snippet that deviates sharply from $\bar{k}_f$ is amplified, even if its semantic match with the analyst report is moderate. The parameter σ_h measures the strength of salience-based amplification, with $\sigma_h = 0$ recovering the baseline. Diagnosticity and salience can coexist because their reference distributions differ: diagnosticity compares recent and historical cross-snippet densities, while salience compares the current snippet with a firm-specific reference embedding.

A.2.6 Confirmation bias

Rabin and Schrag (1999) present a formal model of confirmatory bias: agents who already hold a prior belief misread or over-attend to evidence that confirms it, producing persistent overconfidence. Pouget, Sauvagnat, and Villeneuve (2017) document evidence consistent with confirmatory bias in financial markets, showing that investors are slow to revise opinions when fresh disclosures challenge their existing view.

The confirmation-bias extension amends the logit by introducing an asymmetry based on the analyst's prior stance. Let $s_{a,f} \in \{-1, +1\}$ denote the sign of analyst a's most recent stance on firm f, such as the sign of her most recent forecast revision or recommendation change. Let $\mathrm{tone}(\tilde{k}_{jt,h})$ be the projection of the snippet embedding onto a learned positive-tone direction in semantic space. The extended logit is

$$z_{ijt,h} = \frac{\tilde{q}_{it,h}^{\top}\tilde{k}_{jt,h}}{\tau_h} + \beta_h s_{a,f} \mathrm{tone}(\tilde{k}_{jt,h}). \tag{5}$$

The product $s_{a,f}\mathrm{tone}(\tilde{k}_{jt,h})$ is positive when the snippet's tone agrees with the analyst's prior stance and negative when it disagrees. The coefficient β_h measures the strength of the confirmatory tilt, with $\beta_h = 0$ recovering the symmetric baseline. A positive estimate of β_h would extend the Rabin and Schrag (1999) mechanism to the disclosure-processing stage that precedes a published forecast.

A.3 A combined behaviorally augmented specification

The six extensions above modify the architecture at four different points: the key set, the per-pair logit, the pooling rule, and the training loss. They can therefore be combined into a single

behaviorally augmented specification. The augmented logit is

$$z_{ijt,h} = \underbrace{\frac{\tilde{q}_{it,h}^{\top}\tilde{k}_{jt,h}}{\tau_h}}_{\text{semantic match}} \underbrace{-\delta_h\left(\Delta_{jt}^{\text{age}}\right)^{\alpha_h}}_{\text{fading memory}} + \underbrace{\theta_h \log \frac{p(\tilde{k}_{jt,h} \mid \mathcal{R}_t)}{p(\tilde{k}_{jt,h} \mid \mathcal{B}_t)}}_{\text{diagnosticity}} + \underbrace{\sigma_h \left\|\tilde{k}_{jt,h} - \bar{k}_f\right\|}_{\text{salience}} + \underbrace{\beta_h s_{a,f}\text{tone}(\tilde{k}_{jt,h})}_{\text{confirmation}}. \tag{6}$$

The pooling stage allows a Bayesian prior anchor,

$$e_{t,h} = \lambda_{a,f,h}\mu_{a,f}^{\text{prior}} + (1-\lambda_{a,f,h})\sum_j w_{jt,h}k_{jt}, \tag{7}$$

and the training loss adds sticky inattention,

$$\mathcal{L}_t = \mathcal{L}_t^{\text{pred}} + \kappa_h \left\|e_{t,h} - e_{t-1,h}\right\|^2. \tag{8}$$

The combined specification introduces several behavioral objects: memory parameters (δ_h, α_h), diagnosticity θ_h, salience σ_h, confirmation β_h, prior anchoring $\lambda_{a,f,h}$, and sticky inattention. The sticky-inattention penalty is governed by the loss weight κ_h. Each object has a theoretical interpretation and is identified from a different empirical moment: episode-age decay for (δ_h, α_h), recent-versus-historical embedding density for θ_h, firm-specific contrast for σ_h, prior-stance asymmetry for β_h, within-analyst-firm persistence for $\lambda_{a,f,h}$, and quarter-to-quarter revision smoothness for κ_h. Estimating this specification on the same analyst-disjoint split would turn the precision gap in Section 5 into a decomposition across diagnosticity, salience, confirmation, prior anchoring, and fading memory, rather than leaving it as a purely descriptive attention pattern.

A.4 Forecast-Value Benchmark Horserace

The main text summarizes the forecast-value benchmark comparison; the full horizon-by-horizon horserace is shown here.

A.5 Softmax operator robustness

The main text reports the sparsemax operator because exact zero attention weights define an economically interpretable selected-information set. This appendix repeats the same value and forecast-error validation exercises with softmax attention. Softmax produces dense attention, so the random reference also averages over a broader set of call snippets. The figure should therefore be read as a predictive robustness check rather than as the paper's primary selection test.

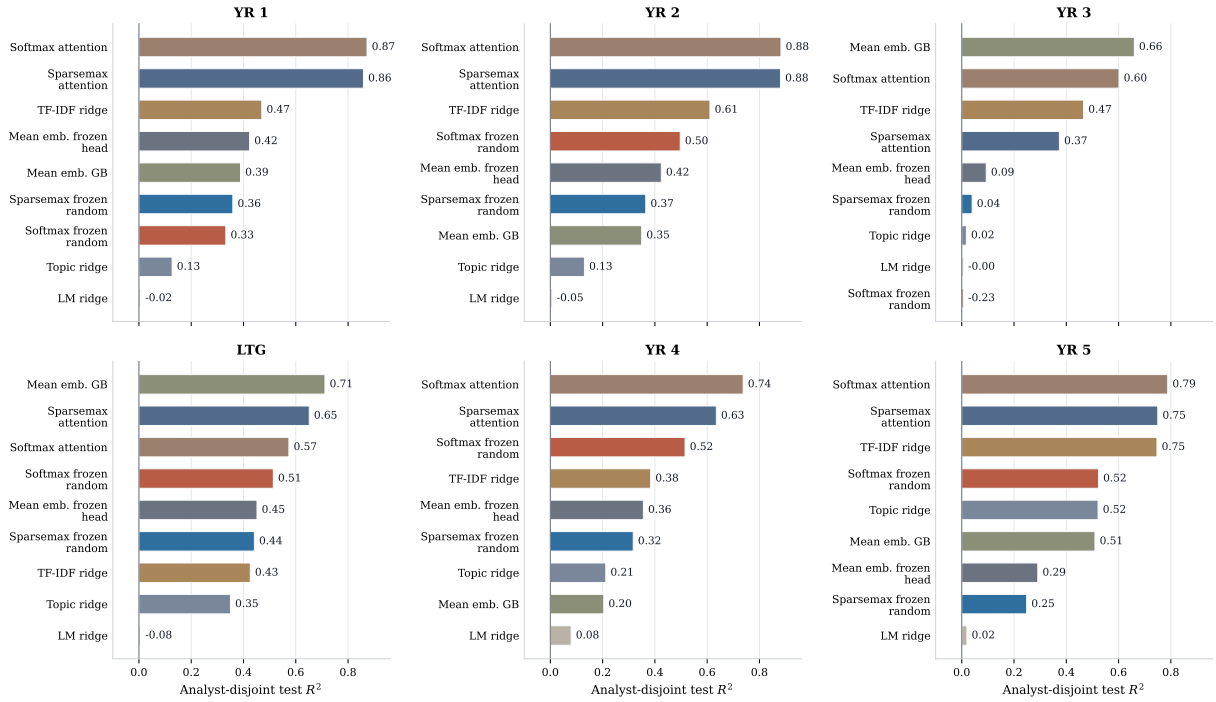


Figure A.1: Forecast-value benchmark comparison. Each panel reports analyst-disjoint test-set R^2 values for one forecast horizon, using forecast values as the outcome. Bars are ordered from highest to lowest within horizon. Sparsemax and softmax attention are full report-query/call-key models trained directly on forecast values; their frozen-random same-k_{eff} references pass randomly selected same-sparsity embeddings through the corresponding value-trained head with the head held fixed. The mean-embedding, TF–IDF, topic-share, and Loughran–McDonald benchmarks are re-estimated with forecast value as the target.

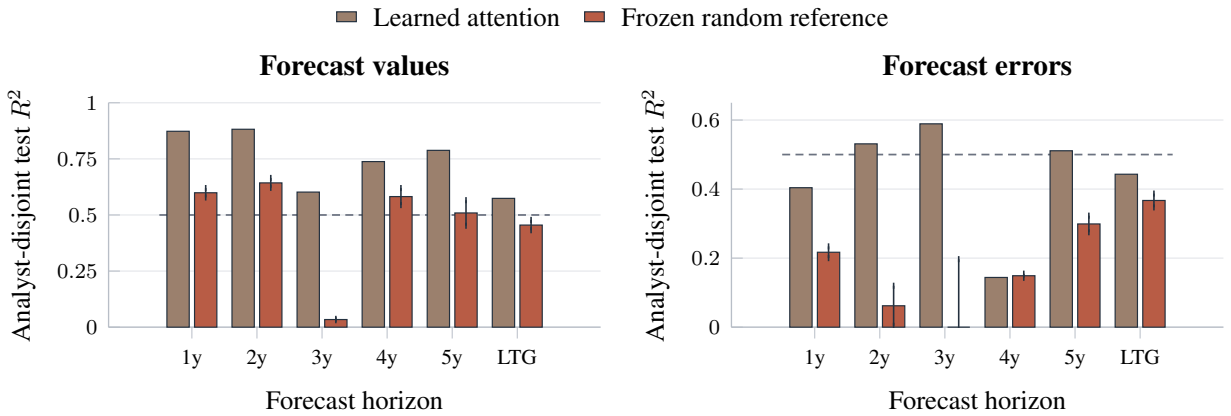
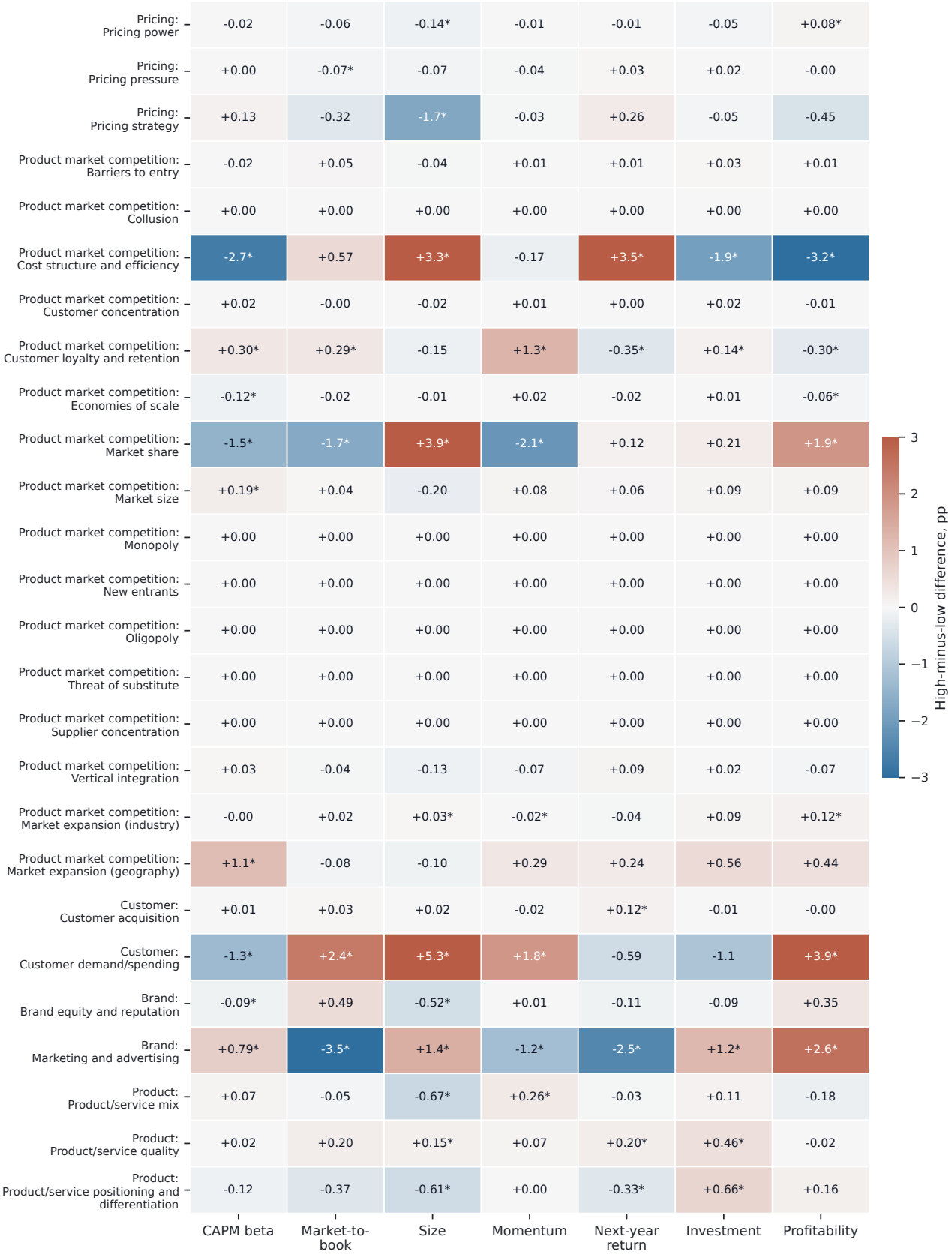
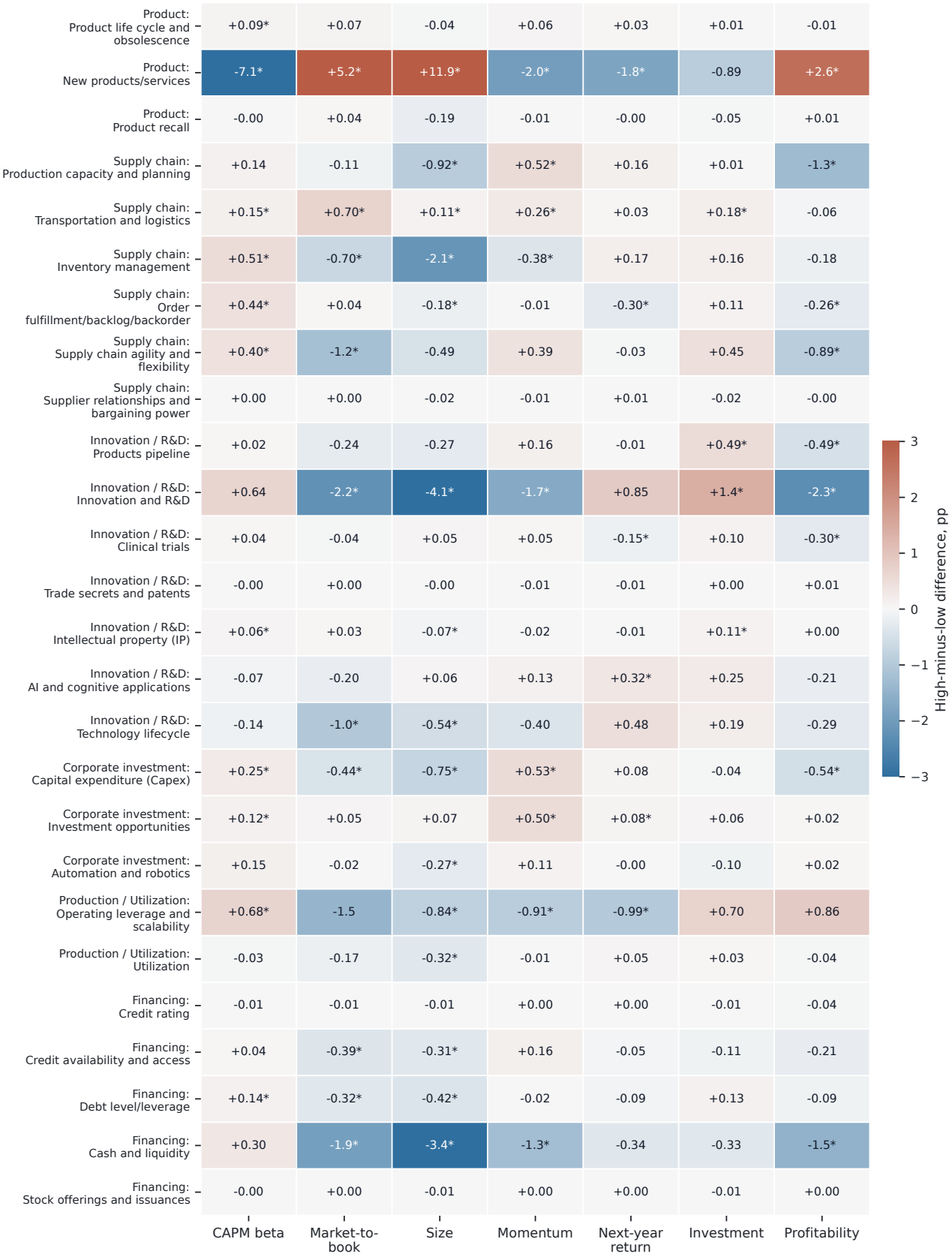


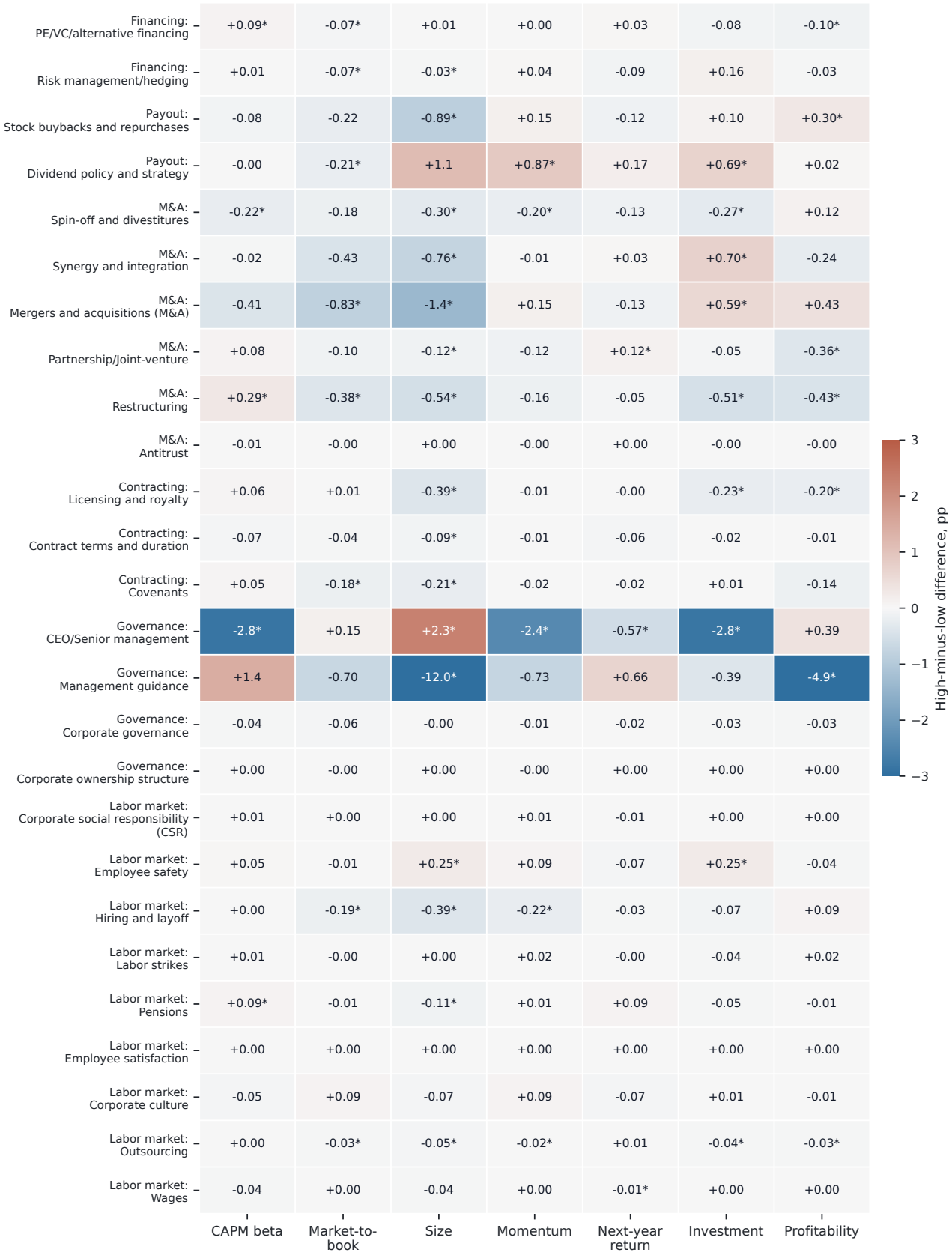
Figure A.2: Softmax attention versus same-k_{eff} frozen-random references. The left panel reports analyst-disjoint test-set R^2 for the softmax attention model trained on forecast values. The right panel reports the same exercise for signed forecast errors. Frozen-random references select the same effective number of call snippets as the corresponding softmax attention model and pass the resulting embedding through the trained prediction head with the head held fixed. Error bars show the standard deviation across random draws.

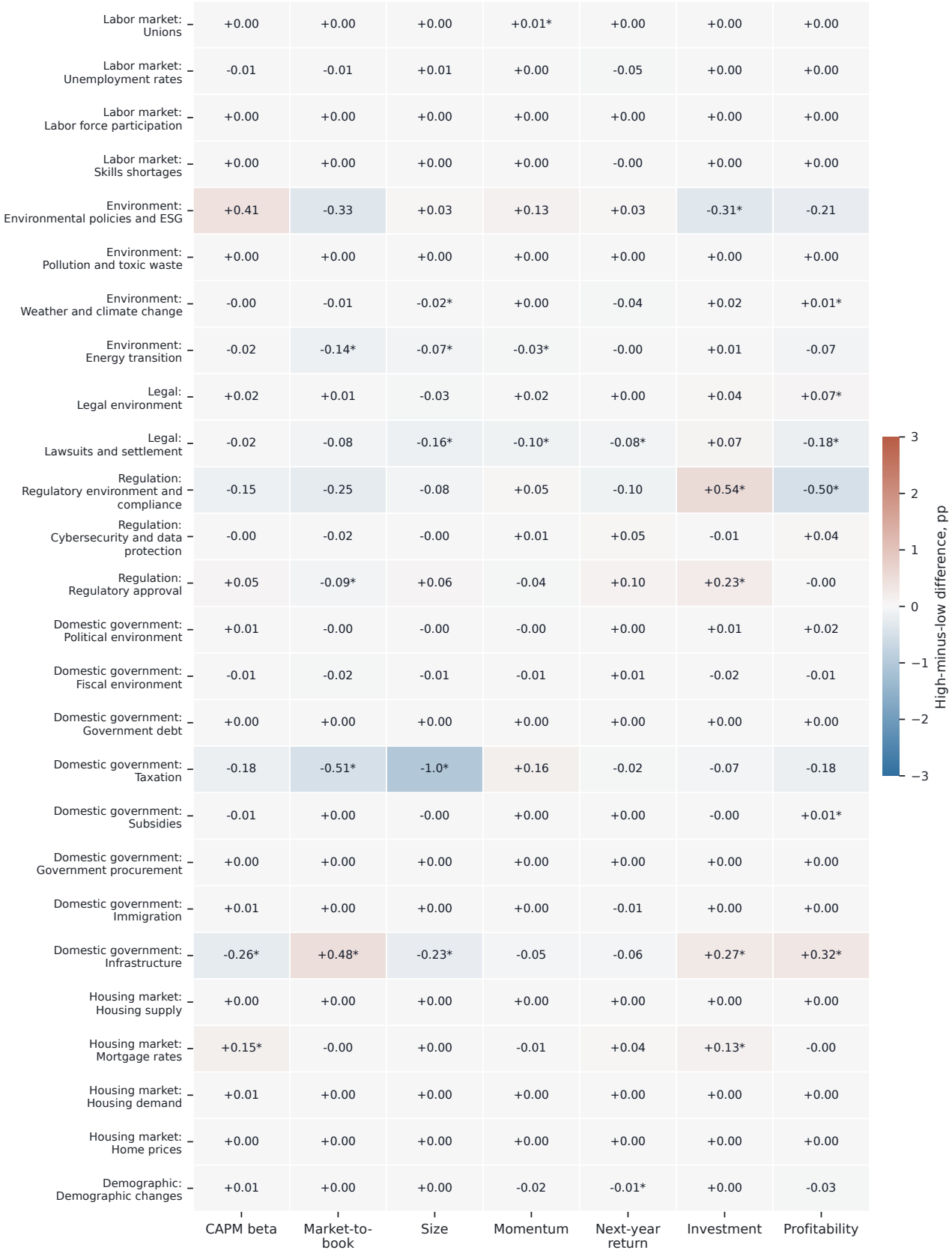
A.6 Complete Forecast-Value Taxonomy Sorts

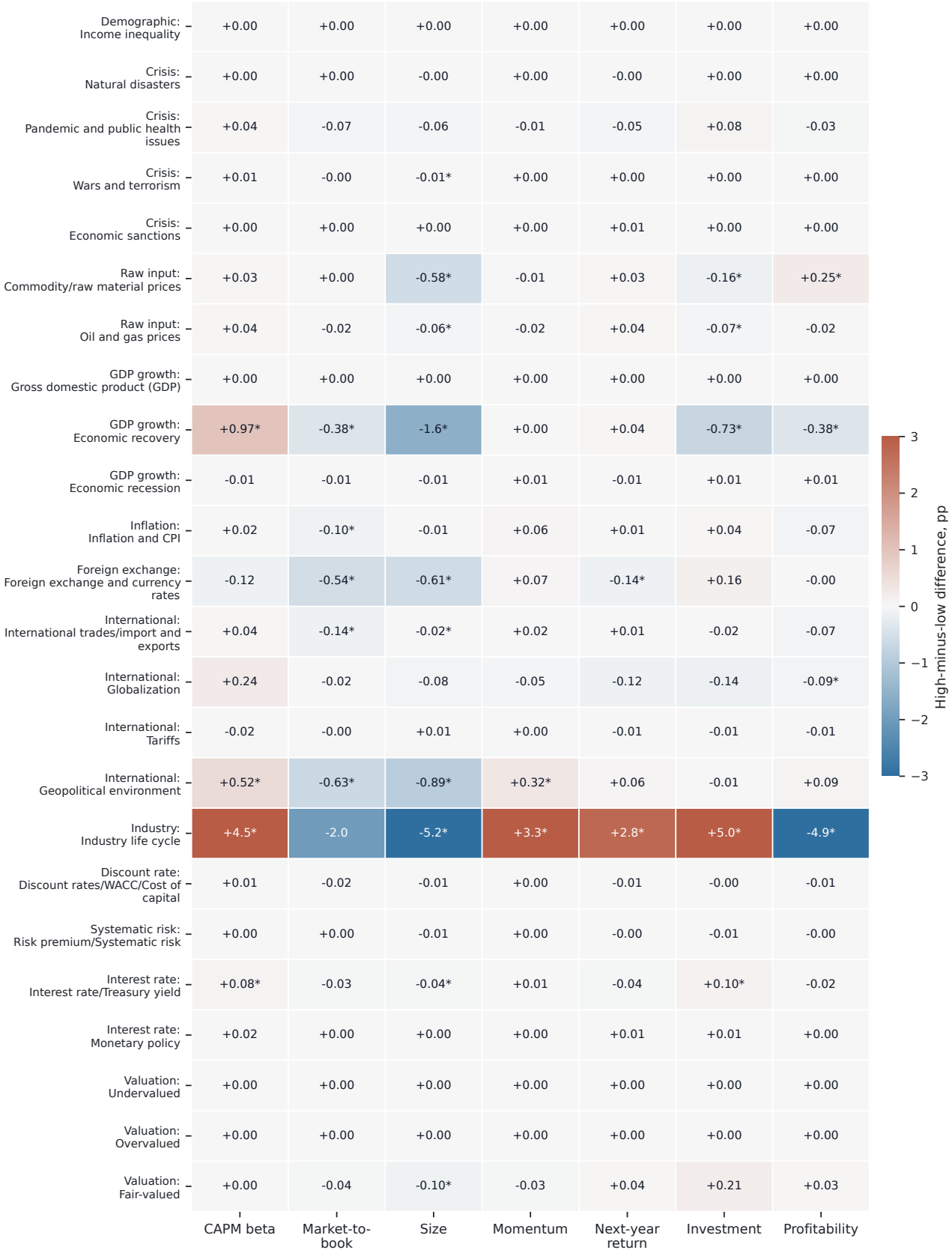
The following five pages repeat Figure 13 for the complete keyword category-label taxonomy. The main-text figure reports only the signal-bearing forecast-value attention labels to keep the comparison readable. The appendix figures show the full taxonomy, including labels with small or statistically indistinguishable sort differences. Columns sort reports into monthly q20/q80 high and low tails on standard characteristics; entries are high-minus-low differences in percentage points after absorbing SIC2 industry-year means. Asterisks mark 95% intervals excluding zero.


Pricing: Pricing power
-0.02 -0.06 -0.14* -0.01 -0.01 -0.05 +0.08*
Pricing: Pricing pressure
+0.00 -0.07* -0.07 -0.04 +0.03 +0.02 -0.00
Pricing: Pricing strategy
+0.13 -0.32 -1.7* -0.03 +0.26 -0.05 -0.45
Product market competition: Barriers to entry
-0.02 +0.05 -0.04 +0.01 +0.01 +0.03 +0.01
Product market competition: Collusion
+0.00 +0.00 +0.00 +0.00 +0.00 +0.00 +0.00
Product market competition: Cost structure and efficiency
-2.7* +0.57 +3.3* -0.17 +3.5* -1.9* -3.2*
Product market competition: Customer concentration
+0.02 -0.00 -0.02 +0.01 +0.00 +0.02 -0.01
Product market competition: Customer loyalty and retention
+0.30* +0.29* -0.15 +1.3* -0.35* +0.14* -0.30*
Product market competition: Economies of scale
-0.12* -0.02 -0.01 +0.02 -0.02 +0.01 -0.06*
Product market competition: Market share
-1.5* -1.7* +3.9* -2.1* +0.12 +0.21 +1.9*
Product market competition: Market size
+0.19* +0.04 -0.20 +0.08 +0.06 +0.09 +0.09
Product market competition: Monopoly
+0.00 +0.00 +0.00 +0.00 +0.00 +0.00 +0.00
Product market competition: New entrants
+0.00 +0.00 +0.00 +0.00 +0.00 +0.00 +0.00
Product market competition: Oligopoly
+0.00 +0.00 +0.00 +0.00 +0.00 +0.00 +0.00
Product market competition: Threat of substitute
+0.00 +0.00 +0.00 +0.00 +0.00 +0.00 +0.00
Product market competition: Supplier concentration
+0.00 +0.00 +0.00 +0.00 +0.00 +0.00 +0.00
Product market competition: Vertical integration
+0.03 -0.04 -0.13 -0.07 +0.09 +0.02 -0.07
Product market competition: Market expansion (industry)
-0.00 +0.02 +0.03* -0.02* -0.04 +0.09 +0.12*
Product market competition: Market expansion (geography)
+1.1* -0.08 -0.10 +0.29 +0.24 +0.56 +0.44
Customer: Customer acquisition
+0.01 +0.03 +0.02 -0.02 +0.12* -0.01 -0.00
Customer: Customer demand/spending
-1.3* +2.4* +5.3* +1.8* -0.59 -1.1 +3.9*
Brand: Brand equity and reputation
-0.09* +0.49 -0.52* +0.01 -0.11 -0.09 +0.35
Brand: Marketing and advertising
+0.79* -3.5* +1.4* -1.2* -2.5* +1.2* +2.6*
Product: Product/service mix
+0.07 -0.05 -0.67* +0.26* -0.03 +0.11 -0.18
Product: Product/service quality
+0.02 +0.20 +0.15* +0.07 +0.20* +0.46* -0.02
Product: Product/service positioning and differentiation
-0.12 -0.37 -0.61* +0.00 -0.33* +0.66* +0.16
CAPM beta
Market-to-book
Size
Momentum
Next-year return
Investment
Profitability
High-minus-low difference, pp
3
2
1
0
−1
−2
−3


Product: Product life cycle and obsolescence
+0.09* +0.07 -0.04 +0.06 +0.03 +0.01 -0.01
Product: New products/services
-7.1* +5.2* +11.9* -2.0* -1.8* -0.89 +2.6*
Product: Product recall
-0.00 +0.04 -0.19 -0.01 -0.00 -0.05 +0.01
Supply chain: Production capacity and planning
+0.14 -0.11 -0.92* +0.52* +0.16 +0.01 -1.3*
Supply chain: Transportation and logistics
+0.15* +0.70* +0.11* +0.26* +0.03 +0.18* -0.06
Supply chain: Inventory management
+0.51* -0.70* -2.1* -0.38* +0.17 +0.16 -0.18
Supply chain: Order fulfillment/backlog/backorder
+0.44* +0.04 -0.18* -0.01 -0.30* +0.11 -0.26*
Supply chain: Supply chain agility and flexibility
+0.40* -1.2* -0.49 +0.39 -0.03 +0.45 -0.89*
Supply chain: Supplier relationships and bargaining power
+0.00 +0.00 -0.02 -0.01 +0.01 -0.02 -0.00
Innovation / R&D: Products pipeline
+0.02 -0.24 -0.27 +0.16 -0.01 +0.49* -0.49*
Innovation / R&D: Innovation and R&D
+0.64 -2.2* -4.1* -1.7* +0.85 +1.4* -2.3*
Innovation / R&D: Clinical trials
+0.04 -0.04 +0.05 +0.05 -0.15* +0.10 -0.30*
Innovation / R&D: Trade secrets and patents
-0.00 +0.00 -0.00 -0.01 -0.01 +0.00 +0.01
Innovation / R&D: Intellectual property (IP)
+0.06* +0.03 -0.07* -0.02 -0.01 +0.11* +0.00
Innovation / R&D: AI and cognitive applications
-0.07 -0.20 +0.06 +0.13 +0.32* +0.25 -0.21
Innovation / R&D: Technology lifecycle
-0.14 -1.0* -0.54* -0.40 +0.48 +0.19 -0.29
Corporate investment: Capital expenditure (Capex)
+0.25* -0.44* -0.75* +0.53* +0.08 -0.04 -0.54*
Corporate investment: Investment opportunities
+0.12* +0.05 +0.07 +0.50* +0.08* +0.06 +0.02
Corporate investment: Automation and robotics
+0.15 -0.02 -0.27* +0.11 -0.00 -0.10 +0.02
Production / Utilization: Operating leverage and scalability
+0.68* -1.5 -0.84* -0.91* -0.99* +0.70 +0.86
Production / Utilization: Utilization
-0.03 -0.17 -0.32* -0.01 +0.05 +0.03 -0.04
Financing: Credit rating
-0.01 -0.01 -0.01 +0.00 +0.00 -0.01 -0.04
Financing: Credit availability and access
+0.04 -0.39* -0.31* +0.16 -0.05 -0.11 -0.21
Financing: Debt level/leverage
+0.14* -0.32* -0.42* -0.02 -0.09 +0.13 -0.09
Financing: Cash and liquidity
+0.30 -1.9* -3.4* -1.3* -0.34 -0.33 -1.5*
Financing: Stock offerings and issuances
-0.00 +0.00 -0.01 +0.00 +0.00 -0.01 +0.00
CAPM beta
Market-to-book
Size
Momentum
Next-year return
Investment
Profitability
High-minus-low difference, pp
3
2
1
0
−1
−2
−3


Financing: PE/VC/alternative financing
Financing: Risk management/hedging
Payout: Stock buybacks and repurchases
Payout: Dividend policy and strategy
M&A: Spin-off and divestitures
M&A: Synergy and integration
M&A: Mergers and acquisitions (M&A)
M&A: Partnership/Joint-venture
M&A: Restructuring
M&A: Antitrust
Contracting: Licensing and royalty
Contracting: Contract terms and duration
Contracting: Covenants
Governance: CEO/Senior management
Governance: Management guidance
Governance: Corporate governance
Governance: Corporate ownership structure
Labor market: Corporate social responsibility (CSR)
Labor market: Employee safety
Labor market: Hiring and layoff
Labor market: Labor strikes
Labor market: Pensions
Labor market: Employee satisfaction
Labor market: Corporate culture
Labor market: Outsourcing
Labor market: Wages
+0.09* -0.07* +0.01 +0.00 +0.03 -0.08 -0.10*
+0.01 -0.07* -0.03* +0.04 -0.09 +0.16 -0.03
-0.08 -0.22 -0.89* +0.15 -0.12 +0.10 +0.30*
-0.00 -0.21* +1.1 +0.87* +0.17 +0.69* +0.02
-0.22* -0.18 -0.30* -0.20* -0.13 -0.27* +0.12
-0.02 -0.43 -0.76* -0.01 +0.03 +0.70* -0.24
-0.41 -0.83* -1.4* +0.15 -0.13 +0.59* +0.43
+0.08 -0.10 -0.12* -0.12 +0.12* -0.05 -0.36*
+0.29* -0.38* -0.54* -0.16 -0.05 -0.51* -0.43*
-0.01 -0.00 +0.00 -0.00 +0.00 -0.00 -0.00
+0.06 +0.01 -0.39* -0.01 -0.00 -0.23* -0.20*
-0.07 -0.04 -0.09* -0.01 -0.06 -0.02 -0.01
+0.05 -0.18* -0.21* -0.02 -0.02 +0.01 -0.14
-2.8* +0.15 +2.3* -2.4* -0.57* -2.8* +0.39
+1.4 -0.70 -12.0* -0.73 +0.66 -0.39 -4.9*
-0.04 -0.06 -0.00 -0.01 -0.02 -0.03 -0.03
+0.00 -0.00 +0.00 -0.00 +0.00 +0.00 +0.00
+0.01 +0.00 +0.00 +0.01 -0.01 +0.00 +0.00
+0.05 -0.01 +0.25* +0.09 -0.07 +0.25* -0.04
+0.00 -0.19* -0.39* -0.22* -0.03 -0.07 +0.09
+0.01 -0.00 +0.00 +0.02 -0.00 -0.04 +0.02
+0.09* -0.01 -0.11* +0.01 +0.09 -0.05 -0.01
+0.00 +0.00 +0.00 +0.00 +0.00 +0.00 +0.00
-0.05 +0.09 -0.07 +0.09 -0.07 +0.01 -0.01
+0.00 -0.03* -0.05* -0.02* +0.01 -0.04* -0.03*
-0.04 +0.00 -0.04 +0.00 -0.01* +0.00 +0.00
CAPM beta
Market-to-book
Size
Momentum
Next-year return
Investment
Profitability
High-minus-low difference, pp
3
2
1
0
−1
−2
−3


Labor market: Unions
+0.00 +0.00 +0.00 +0.01* +0.00 +0.00 +0.00
Labor market: Unemployment rates
-0.01 -0.01 +0.01 +0.00 -0.05 +0.00 +0.00
Labor market: Labor force participation
+0.00 +0.00 +0.00 +0.00 +0.00 +0.00 +0.00
Labor market: Skills shortages
+0.00 +0.00 +0.00 +0.00 -0.00 +0.00 +0.00
Environment: Environmental policies and ESG
+0.41 -0.33 +0.03 +0.13 +0.03 -0.31* -0.21
Environment: Pollution and toxic waste
+0.00 +0.00 +0.00 +0.00 +0.00 +0.00 +0.00
Environment: Weather and climate change
-0.00 -0.01 -0.02* +0.00 -0.04 +0.02 +0.01*
Environment: Energy transition
-0.02 -0.14* -0.07* -0.03* -0.00 +0.01 -0.07
Legal: Legal environment
+0.02 +0.01 -0.03 +0.02 +0.00 +0.04 +0.07*
Legal: Lawsuits and settlement
-0.02 -0.08 -0.16* -0.10* -0.08* +0.07 -0.18*
Regulation: Regulatory environment and compliance
-0.15 -0.25 -0.08 +0.05 -0.10 +0.54* -0.50*
Regulation: Cybersecurity and data protection
-0.00 -0.02 -0.00 +0.01 +0.05 -0.01 +0.04
Regulation: Regulatory approval
+0.05 -0.09* +0.06 -0.04 +0.10 +0.23* -0.00
Domestic government: Political environment
+0.01 -0.00 -0.00 -0.00 +0.00 +0.01 +0.02
Domestic government: Fiscal environment
-0.01 -0.02 -0.01 -0.01 +0.01 -0.02 -0.01
Domestic government: Government debt
+0.00 +0.00 +0.00 +0.00 +0.00 +0.00 +0.00
Domestic government: Taxation
-0.18 -0.51* -1.0* +0.16 -0.02 -0.07 -0.18
Domestic government: Subsidies
-0.01 +0.00 -0.00 +0.00 +0.00 -0.00 +0.01*
Domestic government: Government procurement
+0.00 +0.00 +0.00 +0.00 +0.00 +0.00 +0.00
Domestic government: Immigration
+0.01 +0.00 +0.00 +0.00 -0.01 +0.00 +0.00
Domestic government: Infrastructure
-0.26* +0.48* -0.23* -0.05 -0.06 +0.27* +0.32*
Housing market: Housing supply
+0.00 +0.00 +0.00 +0.00 +0.00 +0.00 +0.00
Housing market: Mortgage rates
+0.15* -0.00 +0.00 -0.01 +0.04 +0.13* -0.00
Housing market: Housing demand
+0.01 +0.00 +0.00 +0.00 +0.00 +0.00 +0.00
Housing market: Home prices
+0.00 +0.00 +0.00 +0.00 +0.00 +0.00 +0.00
Demographic: Demographic changes
+0.01 +0.00 +0.00 -0.02 -0.01* +0.00 -0.03
CAPM beta
Market-to-book
Size
Momentum
Next-year return
Investment
Profitability
High-minus-low difference, pp
3
2
1
0
−1
−2
−3



	CAPM beta	Market-to-book	Size	Momentum	Next-year return	Investment	Profitability
Demographic: Income inequality	+0.00	+0.00	+0.00	+0.00	+0.00	+0.00	+0.00
Crisis: Natural disasters	+0.00	+0.00	-0.00	+0.00	-0.00	+0.00	+0.00
Crisis: Pandemic and public health issues	+0.04	-0.07	-0.06	-0.01	-0.05	+0.08	-0.03
Crisis: Wars and terrorism	+0.01	-0.00	-0.01*	+0.00	+0.00	+0.00	+0.00
Crisis: Economic sanctions	+0.00	+0.00	+0.00	+0.00	+0.01	+0.00	+0.00
Raw input: Commodity/raw material prices	+0.03	+0.00	-0.58*	-0.01	+0.03	-0.16*	+0.25*
Raw input: Oil and gas prices	+0.04	-0.02	-0.06*	-0.02	+0.04	-0.07*	-0.02
GDP growth: Gross domestic product (GDP)	+0.00	+0.00	+0.00	+0.00	+0.00	+0.00	+0.00
GDP growth: Economic recovery	+0.97*	-0.38*	-1.6*	+0.00	+0.04	-0.73*	-0.38*
GDP growth: Economic recession	-0.01	-0.01	-0.01	+0.01	-0.01	+0.01	+0.01
Inflation: Inflation and CPI	+0.02	-0.10*	-0.01	+0.06	+0.01	+0.04	-0.07
Foreign exchange: Foreign exchange and currency rates	-0.12	-0.54*	-0.61*	+0.07	-0.14*	+0.16	-0.00
International: International trades/import and exports	+0.04	-0.14*	-0.02*	+0.02	+0.01	-0.02	-0.07
International: Globalization	+0.24	-0.02	-0.08	-0.05	-0.12	-0.14	-0.09*
International: Tariffs	-0.02	-0.00	+0.01	+0.00	-0.01	-0.01	-0.01
International: Geopolitical environment	+0.52*	-0.63*	-0.89*	+0.32*	+0.06	-0.01	+0.09
Industry: Industry life cycle	+4.5*	-2.0	-5.2*	+3.3*	+2.8*	+5.0*	-4.9*
Discount rate: Discount rates/WACC/Cost of capital	+0.01	-0.02	-0.01	+0.00	-0.01	-0.00	-0.01
Systematic risk: Risk premium/Systematic risk	+0.00	+0.00	-0.01	+0.00	-0.00	-0.01	-0.00
Interest rate: Interest rate/Treasury yield	+0.08*	-0.03	-0.04*	+0.01	-0.04	+0.10*	-0.02
Interest rate: Monetary policy	+0.02	+0.00	+0.00	+0.00	+0.01	+0.01	+0.00
Valuation: Undervalued	+0.00	+0.00	+0.00	+0.00	+0.00	+0.00	+0.00
Valuation: Overvalued	+0.00	+0.00	+0.00	+0.00	+0.00	+0.00	+0.00
Valuation: Fair-valued	+0.00	-0.04	-0.10*	-0.03	+0.04	+0.21	+0.03

A.7 Forecast-Error Snippet Concentration

A useful robustness check is whether the forecast-error signal comes from a single salient call snippet, or effectively from only two snippets, rather than from a broader attended bundle. Table A.1 therefore reports two concentration diagnostics on the analyst-disjoint forecast-error test set. The first is the attention share of the most-attended snippet. The second is the effective number of snippets, $1/\sum_j w_j^2$, which equals two when the attention distribution is as concentrated as two equally weighted snippets. A one- or two-snippet explanation would imply large top-snippet shares and effective counts close to two.

The evidence points against that mechanical interpretation. Across all observation-horizons, the top snippet receives 24.9% of attention on average, only 3.2% of cases put at least half of total attention on one snippet, and only 0.5% have effective attention mass concentrated into two snippets or fewer. The highest-concentration horizon is the two-year forecast, but even there the mean effective count is 5.91 snippets. High-error cases are somewhat more concentrated, but only modestly: the top error quintile has top-snippet attention 2.5 percentage points above the bottom error quintile. Thus the forecast-error model is selective, but its predictions are not driven by a single isolated snippet.

Table A.1: Forecast-error attention is not driven by one or two snippets

Horizon	Top snippet mean (%)	Top snippet $\geq 50\%$ (%)	Effective snippets	Effective ≤ 2 (%)	High–low error top snippet (pp)
YR 1	20.9	0.2	9.51	0.0	3.8
YR 2	32.8	9.2	5.91	1.3	0.8
YR 3	18.8	0.1	10.86	0.0	0.8
YR 4	25.1	2.0	8.30	0.5	7.6
YR 5	26.4	2.0	7.77	0.6	7.1
LTG	20.2	0.0	9.40	0.0	1.9
All horizons	24.9	3.2	8.40	0.5	2.5

Notes: The table uses the analyst-disjoint forecast-error test set. The top-snippet columns report the mean attention share of the most-attended call snippet and the percentage of observation-horizons for which that snippet receives at least half of total attention. Effective snippets are $1/\sum_j w_j^2$, the inverse attention HHI. "Effective ≤ 2" reports the percentage of observation-horizons whose attention distribution is as concentrated as two equally weighted snippets or fewer. The final column is the difference in top-snippet attention between the top and bottom quintiles of realized absolute forecast error within each horizon. If forecast-error prediction were mechanically driven by one or two isolated snippets, these columns would be close to one-snippet concentration; instead, attention remains moderately sparse but clearly multi-snippet.

A.8 Predictable and Residual Components of Forecast Errors

The forecast-error predictions can also be used to decompose realized forecast errors into a systematic, text-predictable component and an idiosyncratic residual component. Let $FE_{i,h}$ denote the realized signed forecast error for observation i and horizon h, and let $\widehat{FE}_{i,h}^{OOS}$ be the analyst-disjoint

out-of-sample prediction from the attention model. For each horizon, I estimate the calibration regression

$$FE_{i,h} = \alpha_h + \beta_h \widehat{FE}_{i,h}^{OOS} + u_{i,h}. \tag{9}$$

The fitted value, $\widehat{FE}_{i,h}^{pred} = \alpha_h + \beta_h \widehat{FE}_{i,h}^{OOS}$, is the predictable mistake: the component of the realized forecast error that is recoverable from the report-query/call-key information environment. The residual $u_{i,h} = FE_{i,h} - \widehat{FE}_{i,h}^{pred}$ is the remaining unpredictable component. This calibration step is useful because it allows the model's raw out-of-sample prediction to be rescaled when the neural head is conservative or over-amplified in a particular horizon.

Table A.2 reports the decomposition. The predictable component is economically meaningful: calibrated predictable shares range from 18.7% for four-year forecasts to 75.4% for three-year forecasts, with most horizons around 37–51%. Thus a substantial part of forecast errors is systematic and text-predictable rather than pure noise. The residual component remains large, especially at the four-year horizon, which is useful conceptually: the attention model identifies systematic mistakes but does not mechanically explain all forecast errors.

Table A.2: Decomposing forecast errors into predictable and residual mistakes

Horizon	Calibration intercept	Calibration slope	Predictable share (%)	Raw model R^2 (%)	Corr.
Panel A: Analyst-disjoint test sample					
YR 1	0.002	1.149	37.0	36.2	0.608
YR 2	0.001	1.203	51.1	49.6	0.715
YR 3	-0.001	2.425	75.4	49.2	0.868
YR 4	0.007	0.933	18.7	18.6	0.432
YR 5	0.009	1.023	47.0	46.5	0.686
LTG	0.747	0.686	45.6	35.4	0.675
Panel B: Full sample					
YR 1	0.005	1.671	18.4	15.3	0.429
YR 2	-0.000	1.693	44.2	36.7	0.665
YR 3	0.005	2.000	50.5	37.8	0.711
YR 4	-0.014	1.324	30.4	28.6	0.551
YR 5	0.006	1.274	19.2	18.1	0.438
LTG	0.153	0.913	53.6	53.1	0.732

Notes: Panel A uses analyst-disjoint out-of-sample predictions from the forecast-error attention model. Panel B repeats the decomposition after evaluating the same trained model on the full estimation sample. For each horizon, realized forecast errors are regressed on predicted forecast errors, $FE_{i,h} = \alpha_h + \beta_h \widehat{FE}_{i,h} + u_{i,h}$. The fitted value is the predictable mistake component, and the residual is the unpredictable component. "Predictable share" is the calibrated regression R^2. "Raw model R^2" is the uncalibrated R^2 from the original prediction. "Corr." is the correlation between realized and predicted forecast errors.

References

John R. Anderson and Lael J. Schooler. Reflections of the environment in memory. *Psychological Science*, 2(6):396–408, 1991.

Susan Athey and Guido W. Imbens. Machine learning methods that economists should know about. *Annual Review of Economics*, 11:685–725, 2019. doi: 10.1146/annurev-economics-080217-053433.

Francesca Bastianello, Paul Décaire, and Marius Günzel. Mental models and financial forecasts. *Marius, Mental Models and Financial Forecasts (October 30, 2024)*, 2024.

J. Aislinn Bohren, Josh Hascher, Alex Imas, Michael Ungeheuer, and Martin Weber. Mental models of information: The role of attention and memory. Working paper, University of Pennsylvania, Yale University, University of Chicago, Aalto University, and University of Mannheim, 2025. URL https://static1.squarespace.com/static/5a1b26c580bd5e0fc6dbd478/t/69325bc5b05b3d5b63d2a43d/1764907975054/BHIUW_MentalModelsInfo.pdf. December 4, 2025.

Pedro Bordalo, Nicola Gennaioli, and Andrei Shleifer. Salience theory of choice under risk. *The Quarterly Journal of Economics*, 127(3):1243–1285, 2012.

Pedro Bordalo, Nicola Gennaioli, and Andrei Shleifer. Salience and asset prices. *American Economic Review*, 103(3):623–628, 2013.

Pedro Bordalo, Nicola Gennaioli, and Andrei Shleifer. Diagnostic expectations and credit cycles. *The Journal of Finance*, 73(1):199–227, 2018.

Pedro Bordalo, Nicola Gennaioli, Rafael La Porta, and Andrei Shleifer. Diagnostic expectations and stock returns. *The Journal of Finance*, 74(6):2839–2874, 2019. doi: 10.1111/jofi.12833.

Pedro Bordalo, Nicola Gennaioli, Yueran Ma, and Andrei Shleifer. Overreaction in macroeconomic expectations. *American Economic Review*, 110(9):2748–2782, 2020.

Pedro Bordalo, Nicola Gennaioli, Rafael La Porta, and Andrei Shleifer. Finance without exotic risk. Technical report, National Bureau of Economic Research, 2024.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning*, pages 2206–2240. PMLR, 2022.

Jean-Philippe Bouchaud, Philipp Krueger, Augustin Landier, and David Thesmar. Sticky expectations and the profitability anomaly. *The Journal of Finance*, 74(2):639–674, 2019.

Leland Bybee. Surveying generative ai's economic expectations. *arXiv preprint arXiv:2305.02823*, 2023.

Hailiang Chen, Byoung-Hyoun Hwang, and Zhuozhen Peng. Why do investors like short-leg securities? evidence from a textual analysis of buy recommendations, 2025. Working paper.

Zhongtian Chen. *Memory and beliefs in financial markets: A machine learning approach*. PhD thesis, University of Pennsylvania, 2025.

Christopher Clayton and Antonio Coppola. Financial regulation and AI: A Faustian bargain? Working Paper 4296, Stanford Graduate School of Business, 2026.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019.

Elliot Epstein, Rose Wang, Jaewon Choi, and Markus Pelger. Attention factors for statistical arbitrage. Working Paper 5588830, SSRN, 2025.

Xavier Gabaix. A sparsity-based model of bounded rationality. *The Quarterly Journal of Economics*, 129(4):1661–1710, 2014.

Xavier Gabaix. Behavioral inattention. In *Handbook of Behavioral Economics: Applications and Foundations 1*, volume 2, pages 261–343. Elsevier, 2019.

Xavier Gabaix, Ralph S. J. Koijen, Robert J. Richmond, and Motohiro Yogo. Asset embeddings. NBER Working Paper 33651, National Bureau of Economic Research, 2025.

Samuel J. Gershman. A unifying probabilistic view of associative learning. *PLoS Computational Biology*, 11(11):e1004567, 2015. doi: 10.1371/journal.pcbi.1004567.

Samuel J Gershman, Ila Fiete, and Kazuki Irie. Key-value memory in the brain. *Neuron*, 113(11): 1694–1707, 2025.

Noah D. Goodman and Michael C. Frank. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818–829, 2016.

Robin Greenwood and Andrei Shleifer. Expectations of returns and expected returns. *The Review of Financial Studies*, 27(3):714–746, 2014.

Thomas L. Griffiths, Falk Lieder, and Noah D. Goodman. Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, 7(2):217–229, 2015.

Zvi Griliches. Economic data issues. In Zvi Griliches and Michael D. Intriligator, editors, *Handbook of Econometrics*, volume 3, pages 1465–1514. Elsevier, 1986.

James J. Heckman. Micro data, heterogeneity, and the evaluation of public policy: Nobel lecture. *Journal of Political Economy*, 109(4):673–748, 2001. doi: 10.1086/322086.

Marc W. Howard and Michael J. Kahana. A distributed representation of temporal context. *Journal of Mathematical Psychology*, 46(3):269–299, 2002.

Daniel Kahneman. *Attention and Effort*. Prentice-Hall, Englewood Cliffs, NJ, 1973.

Daniel Kahneman and Amos Tversky. Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3(3):430–454, 1972.

Shikun Ke. Analysts' belief formation in their own words. *Available at SSRN 5025830*, 2024. doi: 10.2139/ssrn.5025830.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Manish Raghavan. The inversion problem: Why algorithms should infer mental state and not just predict behavior. *Perspectives on Psychological Science*, 19(5):827–838, 2024.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. Matryoshka representation learning. *Advances in Neural Information Processing Systems*, 35: 30233–30249, 2022.

Rafael La Porta. Expectations and the cross-section of stock returns. *The Journal of Finance*, 51 (5):1715–1742, 1996. doi: 10.1111/j.1540-6261.1996.tb05223.x.

Kelvin Law. Good-bye ibes (or not?). *Nanyang Business School Research Paper*, 2022. Research Paper No. 22-09.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Bartosz Maćkowiak and Mirko Wiederholt. Optimal sticky prices under rational inattention. *American Economic Review*, 99(3):769–803, 2009.

Ulrike Malmendier and Stefan Nagel. Depression babies: Do macroeconomic experiences affect risk taking? *The Quarterly Journal of Economics*, 126(1):373–416, 2011.

Ulrike Malmendier and Stefan Nagel. Learning from inflation experiences. *The Quarterly Journal of Economics*, 131(1):53–87, 2016.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Sendhil Mullainathan and Jann Spiess. Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106, 2017. doi: 10.1257/jep.31.2.87.

Stefan Nagel and Zhengyang Xu. Asset pricing with fading memory. *The Review of Financial Studies*, 35(5):2190–2245, 2022.

Sean M. Polyn, Kenneth A. Norman, and Michael J. Kahana. A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review*, 116(1):129–156, 2009.

Sébastien Pouget, Julien Sauvagnat, and Stéphane Villeneuve. A mind is a terrible thing to change: Confirmatory bias in financial markets. *The Review of Financial Studies*, 30(6):2066–2109, 2017.

Matthew Rabin and Joel L. Schrag. First impressions matter: A model of confirmatory bias. *The Quarterly Journal of Economics*, 114(1):37–82, 1999.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Lukas Gruber, Markus Holzleitner, Thomas Adler, David Kreil, Michael K. Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need. In *International Conference on Learning Representations*, 2021.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992, 2019.

Richard Roll. r^2. *The Journal of Finance*, 43(3):541–566, 1988. doi: 10.1111/j.1540-6261.1988.tb04591.x.

Herbert A. Simon. A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1):99–118, 1955.

Herbert A. Simon. Theories of decision-making in economics and behavioral science. *The American Economic Review*, 49(3):253–283, 1959.

Christopher A. Sims. Implications of rational inattention. *Journal of Monetary Economics*, 50(3): 665–690, 2003. doi: 10.1016/S0304-3932(03)00029-1.

Kimberly L. Stachenfeld, Matthew M. Botvinick, and Samuel J. Gershman. The hippocampus as a predictive map. *Nature Neuroscience*, 20(11):1643–1653, 2017.

Joshua B. Tenenbaum, Charles Kemp, Thomas L. Griffiths, and Noah D. Goodman. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022):1279–1285, 2011.

Laura L Veldkamp. *Information choice in macroeconomics and finance*. Princeton University Press, 2011.

Jessica A. Wachter and Michael J. Kahana. A retrieved-context theory of financial decisions. *The Quarterly Journal of Economics*, 139(2):1095–1147, 2024.