A Financial Brain Scan of the LLM *

Hui Chen[†], Antoine Didisheim[‡], Mo Pourmohammadi[§],
Luciano Somoza[¶], Hanqing Tian[‖]

August 13, 2026

Abstract

We use sparse autoencoders (SAEs) to "brain scan" large language models, mapping internal activations to interpretable concepts and enabling targeted "steering" of specific features. We show that prompting models to adopt specific preferences activates many unintended concepts, complicating causal inference in agent simulations. In contrast, SAEs permit smooth, more isolated steering of target concepts. Furthermore, our scans reveal a stark disconnect between the model's internal feature contributions and its self-generated explanations, showing that these explanations are unreliable guides to its reasoning. Finally, SAEs provide a lightweight intervention for existing models, producing interpretable embeddings and allowing researchers to attenuate specific economic biases.

*We thank participants at the Toronto Asset Pricing and Investments Workshop 2026 and the Stanford Institute for Theoretical Economics (SITE) 2026, as well as seminar participants at Oxford University and the University of Technology Sydney for their valuable comments. Any remaining errors are our own. The authors have nothing to disclose.

[†]MIT Sloan and NBER. Email: huichen@mit.edu
[‡]University of Melbourne. Email: antoine.didisheim@unimelb.edu.au
[§]Yale SOM. Email: mo.pourmohammadi@yale.edu
[¶]ESSEC Business School. Email: somoza@essec.edu
[‖]University of Melbourne. Email: hanqing.tian1@unimelb.edu.au

I. Introduction

Large language models (LLMs) are increasingly used in financial analysis, research workflows, and investment decisions (Cheng et al., 2024). Their appeal is clear: they can process and summarize large volumes of unstructured text. Among the challenges these models present, two are central to our analysis. First, their internal representations are dense and difficult to interpret, making it hard to identify which concepts drive their outputs (Ludwig et al., 2025). Second, their outputs can reflect latent biases, including systematic differences associated with demographic characteristics (Fedyk et al., 2024). We address these challenges by making the model's internal representations interpretable and controllable.

We apply sparse autoencoders (SAEs), an emerging interpretability technique from computer science (see, e.g., Cunningham et al., 2023; Gao, 2024; Shi, 2025), to economic applications. An SAE is a small auxiliary network attached to an already trained LLM. It transforms the model's dense internal state into a larger, sparse representation in which only a small subset of features activates for any given input. These features can often be associated with interpretable concepts. Researchers can therefore observe which concepts activate as the model processes an input and intervene by increasing or suppressing selected activations without retraining the LLM or modifying its underlying weights.

Off-the-shelf SAEs are general-purpose, so many of their features capture concepts that are unrelated to economics and finance. We therefore develop a finance-specific labeling and selection procedure that maps 7,278 SAE features to 46 economic topics organized into five high-level themes. The resulting representation allows us to use the SAE as a "brain scan" of the model's economic reasoning and to target specific economic concepts through steering.

Beyond interpreting the model, the labeled representation allows us to compare two methods of constructing economic agents: prompt steering and SAE steering. Prompt steering induces a trait, preference, or bias by describing it in the input, for example, "You are a highly risk-averse investor." We use risk aversion as a concrete application and show that this prompting approach fails to isolate the targeted concept. Asking the model to become more risk-averse simultaneously activates many unintended concepts, making it difficult to attribute behavioral changes to the intended preference. Prompt steering also produces abrupt changes rather than smooth, calibrated adjustments in behavior. By contrast, direct steering of the SAE's *Risk Aversion* feature changes the targeted concept continuously while producing substantially smaller changes in other measured features. These properties provide a more credible basis for causal analysis in economic agent simulations.

Constructing and Validating the Brain Scan. The standard approach to labeling SAE features is bottom-up. Researchers pass a general-purpose text corpus through the model, identify the passages that most strongly activate each feature, and ask another LLM to describe their shared theme. Because both the SAE and the corpus are general-purpose, the resulting labels span subjects ranging from cooking recipes to sports scores, and most are of little use for financial research. Our first contribution is a finance-specific labeling and selection pipeline. We define 46 granular economic topics in advance, including *Leverage*, *Monetary Policy*, and *Risk Aversion*. We organize these topics into five broad themes: *Corporate Finance*, *Investment & Portfolio Management*, *Risk Management & Derivatives*, *Financial Markets & Instruments*, and *Macro-Finance & Economics*. We generate representative texts for each topic and assign each SAE feature to the topic for which it activates most strongly and selectively. This procedure both labels the features and isolates the subset relevant to our analysis, reducing the feature space from 131,072 features to 7,278 economically relevant ones.

To assess both our labeling procedure and the broader usefulness of SAEs for financial analysis, we conduct two initial validation exercises. First, a controlled word-cloud test validates our topic labels. We give the model the same neutral prompt, "Give me 3 finance/economic concepts without explanation," while steering its internal state toward each of our 46 topics in turn. With the prompt held fixed, the model consistently generates vocabulary associated with the targeted topic. For example, steering toward *Derivatives* produces terms such as "option," "call," and "straddle," whereas steering toward *Governance* produces "board," "succession," and "activist." The alignment between the steered topic and the generated vocabulary provides direct validation of our topic assignments.

Second, we deploy our "brain scan" to observe the model's internal reasoning during a complex task: the formation of macroeconomic expectations. Following Bybee (2023), we prompt the LLM to read news articles and form expectations about fourteen macroeconomic variables. For thirteen of the fourteen variables, the correlations between the LLM and human survey expectations are broadly in line with the baseline results reported by Bybee (2023). This establishes that the SAE-augmented model continues to produce economically meaningful expectations. When we aggregate the gradient-based contributions into our five high-level themes, the model's internal contributions vary markedly across forecast targets. For instance, the S&P 500 forecast relies overwhelmingly on *Financial Markets & Instruments*, whereas the Consumer Price Index forecast relies heavily on *Macro-Finance & Economics*, particularly *Inflation*. The model processes the same news in both cases. The only difference is the variable it is asked to forecast. Its internal contributions therefore shift with the economic question, providing a second validation of our topic labels in a more

complex setting. When we ask the model to explain its forecasts, however, its self-generated explanations frequently diverge from the contributions revealed by our brain scan. This discrepancy shows that the model's verbal explanations are an unreliable guide to the internal concepts that drive its forecasts.

Steering & Agent Simulation. Researchers increasingly use LLMs to simulate economic agents (e.g., Horton, 2023). A growing literature shows that these models can reproduce decision-making patterns observed across different human populations and economic settings (e.g., Fedyk et al., 2024; Aher et al., 2023; Argyle et al., 2023). Researchers can therefore use a single LLM to construct heterogeneous simulated agents by assigning different traits through the prompt. For example, a prompt may assign the agent a demographic profile ("You are a female investor..."), a psychological preference ("You are highly risk-averse..."), or a cognitive bias. The underlying model remains fixed, and differences in behavior across simulated agents are attributed to the traits assigned through these prompts.

This approach relies on the assumption that changing the prompt alters only the assigned trait. We show that this assumption fails. Prompting the model to adopt a particular trait changes not only the targeted concept but also many unrelated concepts. Behavioral differences across prompts therefore cannot be attributed solely to the intended preference, compromising causal interpretation in prompt-steered agent simulations.

We demonstrate this limitation in a controlled simulation in which the LLM chooses between two investment opportunities. We increase the agent's risk aversion using two methods. Prompt steering adds explicit instructions that tell the model to become more risk-averse, whereas SAE steering directly increases the activation of a selected *Risk Aversion* feature. Our brain scan reveals that both methods increase the targeted concept, but their broader effects differ substantially. SAE steering directly modifies the selected feature, while prompt steering also changes many unrelated concepts, some by magnitudes comparable to the change in risk aversion. Prompt-induced changes in behavior therefore combine the effect of risk aversion with changes in other dimensions of the model's reasoning.

This contamination has direct consequences for the model's economic reasoning. To demonstrate this, we present the LLM with a choice between two projects with identical risk but different net present values (NPVs). Because the higher-NPV project strictly dominates, the correct choice should be invariant to the agent's level of risk aversion. Under SAE steering, the model's decision remains unchanged across all steering strengths we test, and it consistently selects the higher-NPV project. Under prompt steering, however, the model eventually reverses its decision and selects the dominated, lower-NPV project as the instruc-

tion to become more risk-averse intensifies. Our brain scan shows that this reversal coincides with changes in unrelated concepts, including a reduction in the activation associated with *Capital Budgeting*. We obtain the same general pattern across alternative prompts and decision environments. Prompt steering changes concepts unrelated to the intended preference and can overturn otherwise correct financial decisions, whereas SAE steering preserves the correct response throughout the range of interventions we examine.

A second limitation of prompt-based steering is its inability to produce smooth, continuous changes in agent behavior. To demonstrate this, we present the LLM with a binary choice between a safe and a risky investment. Following Chen et al. (2024), we examine the next-token probability that the model assigns to each decision. Rather than observing only the model's final choice, these probabilities reveal how strongly the model favors one option over the other. The probability assigned to the risky asset can range from nearly 0%, when the model strongly favors the safe asset, to nearly 100%, when it strongly favors the risky asset, with intermediate values capturing less decisive preferences. As we increase the activation of the SAE's *Risk Aversion* feature, this probability changes continuously, allowing us to move the model gradually from the risky asset toward the safe asset. Prompt steering does not provide the same granularity. Even when we vary the instruction incrementally by asking the model to become "1%," "2%," or "3% more risk-averse," its decision probability changes through abrupt, discrete jumps rather than gradual adjustments.

This inability to produce fine-grained, continuous changes creates an important problem for experimental design. Researchers need to calibrate the strength of a preference and apply the same intervention across different decision settings. SAE steering provides such a numerical scale. A researcher can calibrate a steering coefficient using one representative prompt and then apply the same coefficient, without recalibration, to a different but related prompt. The coefficient therefore provides a common, quantifiable scale across prompts. We show that this calibration produces similar behavioral responses across alternative prompt wordings. Prompt steering lacks this portability. A textual instruction that produces a small change in one setting may generate an abrupt change, or no change, in another. Moreover, an instruction such as "10% more risk-averse" has no stable correspondence to either the model's internal representation or its behavior. Prompt-based preferences are therefore difficult to standardize across simulation environments.

Steering and Bias Correction. Existing research documents several biases in LLM outputs, ranging from systematic differences associated with demographic characteristics (Fedyk et al., 2024) to a general optimism bias associated with post-training alignment (Liu and Kirshner, 2024). LLMs can also rely on spurious or internally inconsistent economic rea-

soning. These problems limit their usefulness for financial analysis in different ways: latent biases can systematically tilt predictions, while flawed reasoning can produce economically incoherent forecasts. We show that SAEs provide an interpretable way to identify the internal concepts underlying these errors and intervene on them directly. We demonstrate this application of SAE steering by correcting a reasoning failure in macroeconomic forecasting and optimism bias in news-based return prediction.

A compelling example of spurious reasoning emerges from our earlier analysis of LLM-generated macroeconomic expectations. For thirteen of the fourteen variables, the correlations between LLM and survey expectations are broadly in line with the baseline results of Bybee (2023). The notable exception is the revision in the 3-month Treasury bill rate. Plotting the time series reveals that this discrepancy is concentrated during the 2008 financial crisis, when the LLM predicts an increase in the rate while professional forecasters predict a decline. When asked to explain its forecast, the model invokes a "flight-to-safety" narrative to predict an increase in the Treasury bill rate. This reasoning fails on two counts. First, a flight to safety increases demand for Treasury bills, raising their prices and lowering their yields. The mechanism therefore implies the opposite of the model's prediction. Second, the model overlooks the central bank's response to the crisis, including the rate cuts used to stimulate the economy. We address this omission by increasing the activation of the SAE feature most strongly associated with *Quantitative Easing*. This targeted intervention shifts the model's reasoning toward a coherent monetary-policy explanation in which the central bank lowers rates in response to the crisis, causing the Treasury bill rate to fall. Steering this single, interpretable feature removes the anomalous rate spike during 2008 and raises the model's correlation with survey revisions to a level comparable to the baseline reported by Bybee (2023).

We next turn to news-based return prediction to examine whether SAE steering can correct the model's optimism bias. Following Lopez-Lira and Tang (2023) and Chen et al. (2024), we ask the LLM to classify overnight Reuters alerts for individual stocks as positive or negative. Increasing the activation of an SAE feature associated with positive sentiment monotonically raises the share of alerts classified as positive. Decreasing its activation has the opposite effect and attenuates the model's baseline tendency toward positive classifications. Importantly, downward steering does not simply flip predictions at random. Instead, it makes the model more selective: fewer alerts are classified as positive, but those that remain in the positive group are followed by larger positive returns. This targeted correction improves the performance of the resulting long-short portfolio, raising its annualized Sharpe ratio from 4.39 for the unsteered model to 4.71 under downward SAE steering. By contrast, standard prompt engineering performs poorly. Explicitly instructing the model to become

"less optimistic" produces abrupt changes in the classification rate and Sharpe ratios well below the unsteered baseline. These results show that SAE steering can attenuate a specific model bias without eliminating the economic information contained in its classifications.

Interpretable Embeddings. An emerging literature uses LLM embeddings, rather than generated classifications, as inputs to downstream predictive models (Chen et al., 2022; Didisheim et al., 2026; Zhang and Zhou, 2026). This approach retains the information contained in the model's internal representations and is generally more tractable and stable than prompt-based classification. Its main limitation is interpretability. Standard embeddings are dense vectors in which economic concepts are distributed across many dimensions, making it difficult to identify the concepts responsible for a prediction. SAEs address this limitation by transforming the model's dense representations into sparse activations that can be mapped to economically meaningful concepts. They therefore retain the tractability of embedding-based methods while allowing us to identify the economic content underlying their predictions.

We next examine whether this interpretability comes at the cost of predictive performance. Following the return-prediction framework of Chen et al. (2022), we use embeddings extracted from firm-level news to predict subsequent stock returns. We restrict the SAE representation to the 7,278 features that our labeling procedure maps to economic topics and compare its performance with that of standard dense embeddings extracted from the same LLM. The results show that interpretability does not reduce predictive content. The long-short portfolio constructed using all labeled SAE features achieves an annualized Sharpe ratio of 5.22, compared with 3.67 for the dense embedding benchmark. Even a representation containing only the 50 most predictive labeled features produces a Sharpe ratio of 3.84, slightly exceeding the benchmark.

The SAE representation also reveals where this predictive content originates. Among the ten most predictive features, *Macro-Finance & Economics* and *Financial Markets & Instruments* each account for approximately 40%. These broad, market-wide concepts therefore contain the strongest individual signals. As we include additional features, the composition shifts toward *Corporate Finance*, which becomes the largest theme once the representation contains approximately 1,000 features. This pattern suggests that a small number of macroeconomic and market-wide concepts provide broad predictive signals, while predictive content related to corporate finance is distributed across a larger set of features capturing narrower, firm-specific information.

Related Literature. Our paper contributes to four strands of literature. The first uses LLMs as empirical tools in finance and economics, including applications to news-based return prediction (Lopez-Lira and Tang, 2023; Chen et al., 2022), financial-statement analysis (Kim et al., 2024a), macroeconomic expectations (Bybee, 2023), investor adoption (Cheng et al., 2024), and economic research methodology (Korinek, 2023; Ludwig et al., 2025). Most of this work takes the model's generated output or embedding as the empirical object. We instead examine the model's internal representations to identify which economic concepts contribute to a given answer.

A second strand uses LLMs to simulate economic agents whose characteristics and roles are assigned through prompts. Applications range from synthetic survey respondents and generative agents (Horton, 2023; Aher et al., 2023; Argyle et al., 2023; Park et al., 2023) to decision makers assigned particular financial advisory roles, risk preferences, or other economic preferences (Fieberg et al., 2025; Liu et al., 2025; Kim et al., 2024b; Cook et al., 2026). Our results identify an important limitation of this approach. A prompt intended to change one trait also activates many unrelated concepts, so prompt-assigned personas do not isolate the intended treatment. Observed behavioral changes may therefore reflect both the assigned trait and collateral changes in other dimensions of the model's reasoning.

A third strand documents biases and fragility in LLM behavior, including demographic and perception biases, forecast optimism, and look-ahead or memorization artifacts (Fedyk et al., 2024; Gallegos et al., 2024; Liu and Kirshner, 2024; Sarkar and Vafa, 2024; Lopez-Lira et al., 2025). We complement this literature by identifying internal concepts associated with these failures and testing whether intervening on those concepts directly can correct the resulting predictions and reasoning.

A fourth strand studies text-based and machine-learning approaches to asset pricing. This literature develops measures based on media sentiment and textual analysis (Tetlock, 2007; Loughran and McDonald, 2011; Engelberg and Parsons, 2011), as well as high-dimensional machine-learning and factor models (Gu et al., 2020; Giglio et al., 2022; Kelly et al., 2024; Gu et al., 2021). More recent work uses dense LLM embeddings of financial text as inputs to predictive models (Chen et al., 2022; Didisheim et al., 2026; Zhang and Zhou, 2026). These embeddings contain substantial predictive information but do not reveal which economic concepts drive their predictions. Our SAE representation maps this predictive content into a sparse set of economically labeled features, allowing us to retain the advantages of embedding-based methods while interpreting their predictions. Methodologically, we build on sparse autoencoders developed in the mechanistic-interpretability literature (Cunningham et al., 2023; Gao, 2024; Lieberum et al., 2024; Shi, 2025) and adapt them to economic

applications through our finance-specific labeling procedure.

The remainder of the paper is organized as follows. Section II introduces the SAE methodology, presents the finance-specific labeling procedure, and validates the resulting labels through a steering exercise. Section III applies the brain scan to the formation of macroeconomic expectations and decomposes the model's forecasts into interpretable economic concepts. Section IV compares SAE and prompt steering in economic agent simulations using a risk-aversion case study. Section V applies SAE steering to two settings: a Treasury bill forecasting error during the 2008 financial crisis and optimism bias in news-based return prediction. Section VI evaluates whether interpretable SAE embeddings retain predictive content in a firm-level news return-prediction exercise. Section VII concludes.

II. Methodology

LLMs process text through a sequence of transformer layers. At each token position, the model maintains a vector representation known as the *residual stream*, which serves as a running representation of the information processed by the model and is updated by each successive layer (left panel of Figure 1). During pretraining, the model learns these representations by predicting the next token. This architecture is highly effective for prediction, but it does not require the resulting representations to be interpretable. The residual stream is *dense*: for a given token, most of its thousands of coordinates take nonzero values, so information is spread throughout the vector rather than concentrated in a small number of active entries. These coordinates also lack simple individual interpretations. Economic concepts such as "inflation" or "optimism" are represented by patterns involving many coordinates, while any given coordinate may contribute to the representation of several concepts. Consequently, inspecting individual dimensions of the residual stream does not reveal which concepts contribute to the model's output.

SAEs offer a straightforward tool for untangling this dense representation.[1] At its core, an SAE is a separate neural network trained to reconstruct the residual-stream vectors produced by an already trained LLM. It employs an encoder-decoder architecture: the encoder takes a dense residual-stream vector as input and projects it into a higher-dimensional intermediate layer, while the decoder maps this representation back to the original dimensions. As illustrated in the right panel of Figure 1, the residual stream at a selected transformer layer is passed through the SAE, expanded into a wider intermediate representation, and then reconstructed before being returned to the LLM.

[1] Appendix A provides formal definitions of the SAE architecture and our concept-steering methodology.

The SAE's training objective balances reconstruction accuracy and sparsity. The reconstruction objective requires the output vector to closely match the original residual stream, thereby preserving the information passed to subsequent transformer layers. The sparsity objective penalizes the simultaneous activation of many features, encouraging the SAE to represent each input using only a small number of nonzero values. The intermediate layer is substantially wider than the original residual stream, giving the SAE a large set of potential features from which to select this small active subset. Rather than encoding many concepts through overlapping patterns in a smaller dense vector, the SAE can represent them using more distinct features that activate only when relevant. This combination of greater dimensionality and sparse activation makes the resulting representation easier to interpret.

A key practical strength of SAEs is that they provide a lightweight addition to an existing LLM. The underlying language model remains unchanged, making an SAE substantially less expensive to train than the model itself. During SAE training, the LLM's parameters remain fixed. A text corpus is passed through the frozen model, residual-stream activations are collected at a selected layer, and the SAE is trained on those activations.

Once trained, an SAE makes the model's internal state both interpretable and controllable. First, it acts as a "brain scan" of the model's reasoning. By monitoring active features and their gradients, we can identify the concepts that contribute to the model's next-token predictions and quantify their relative importance. Second, the SAE enables *concept steering*. By amplifying or suppressing selected features during text generation, we can directly shift the model's internal focus toward or away from a targeted concept. These two capabilities allow us both to observe the economic concepts underlying the model's output and to intervene on them directly.

Throughout the paper, we use the pretrained Gemma Scope SAEs released by Google DeepMind for *Gemma-2-9B-IT* (Lieberum et al., 2024). This setup combines a powerful instruction-tuned model with the computational efficiency required for extensive experimentation. Our subsequent empirical exercises show that it reproduces or exceeds established performance benchmarks in financial applications. The central limitation of the off-the-shelf SAE is therefore not performance, but domain relevance. Its 131,072 features form a large, general-purpose dictionary spanning concepts from finance to cooking, sports, and everyday conversation. Only a small fraction directly captures the economic mechanisms relevant to our analysis. Our first contribution transforms this general-purpose representation into an economic one through a finance-specific procedure that labels and selects the features associated with economically meaningful concepts.

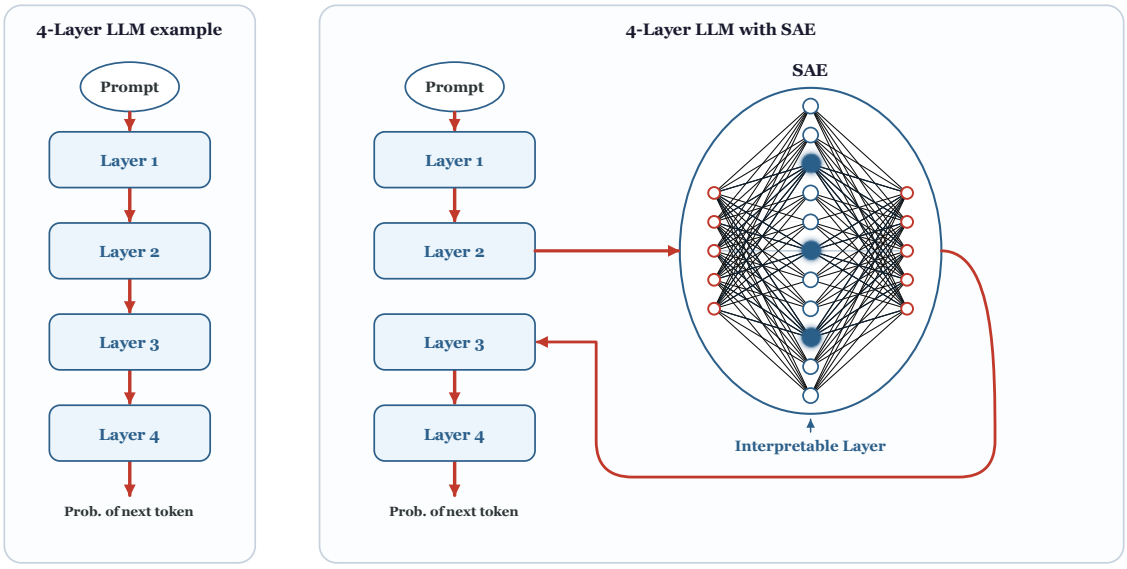


Figure 1. Sparse Autoencoder Representation of the Residual Stream.
The left panel shows a stylized four-layer language model. A prompt passes through successive transformer layers to produce a probability distribution over the next token. The red arrows trace the *residual stream*, the running hidden representation updated and passed from one layer to the next. The right panel shows the same model with an SAE inserted between Layers 2 and 3. The encoder maps the dense residual-stream vector into a substantially wider feature layer. Only a small number of these features are active for a given input, as shown in dark blue. The decoder then reconstructs the vector in its original dimensions and passes it to Layer 3.

A. Finance-specific Feature Labeling and Selection

The features learned by an SAE do not come with economically meaningful labels. The standard labeling procedure is bottom-up: researchers collect the text passages that most strongly activate each feature and ask a separate LLM to summarize their common theme. This approach produces useful general-purpose descriptions, but it is poorly suited to answering specific economic questions. Because the activating passages are drawn from generic text corpora, many of the resulting labels describe everyday subjects such as recipes, sports, or casual conversation rather than the economic concepts needed for financial and macroeconomic analysis.

We address this limitation with a top-down procedure that jointly labels and selects economically relevant SAE features. This procedure reverses the standard approach: we begin with a predefined taxonomy of economic concepts and identify the features most closely associated with each concept. Our taxonomy contains 46 economic topics grouped into five broad categories: *Corporate Finance*, *Investment & Portfolio Management*, *Risk Management & Derivatives*, *Financial Markets & Instruments*, and *Macro-Finance & Economics*.

Figure 2 presents the full taxonomy.

Next, for each of the 46 topics, we prompt an LLM to generate synthetic financial articles centered on that topic.[2] We process these articles through the SAE-augmented model and record the activation of each SAE feature across topics. A feature receives a high assignment score for a topic when it activates strongly on articles about that topic and less strongly on articles about the other topics.[3] We also measure how concentrated each feature's activation is across topics and discard features with diffuse activation patterns. For each topic, we rank the assigned features by their assignment scores and retain the smallest set of top-ranked features that accounts for a fixed share of the topic's total assignment signal. The resulting set, $\mathcal{F}_i$, contains features that activate strongly and selectively for topic i, while excluding weak or broadly activated features.

Across the 46 topics, this procedure identifies 7,278 of the SAE's 131,072 features, or approximately 5.6%, as economically relevant. The procedure therefore performs labeling and dimensionality reduction in a single step. Importantly, the entire procedure relies only on synthetic topic-specific articles. Neither the empirical news sample nor realized stock returns enter the labeling or selection process, keeping the resulting feature set independent of the outcomes examined in our empirical applications.

B. Validating Labels

To validate that the features we identified genuinely capture their assigned economic topics and can be used to effectively steer the model, we run a simple control test. We use a neutral prompt asking the LLM to generate concepts freely, while artificially steering its internal state toward a specific topic.

```
Give me 3 finance/economic concepts without explanation.
```

Prompt 1. Word-Cloud Probe
This neutral prompt elicits a short completion from the LLM while its residual stream is steered along the features assigned to a given topic.

For each identified topic, we steer its associated features to their maximum activation value and ask the model to complete Prompt 1. We repeat this generation three times per topic, pool all the generated words, and render their frequencies into a word cloud.

[2]Prompt 9 in Appendix B provides the exact prompt.
[3]Appendix A.C provides the formal assignment and selection procedure.

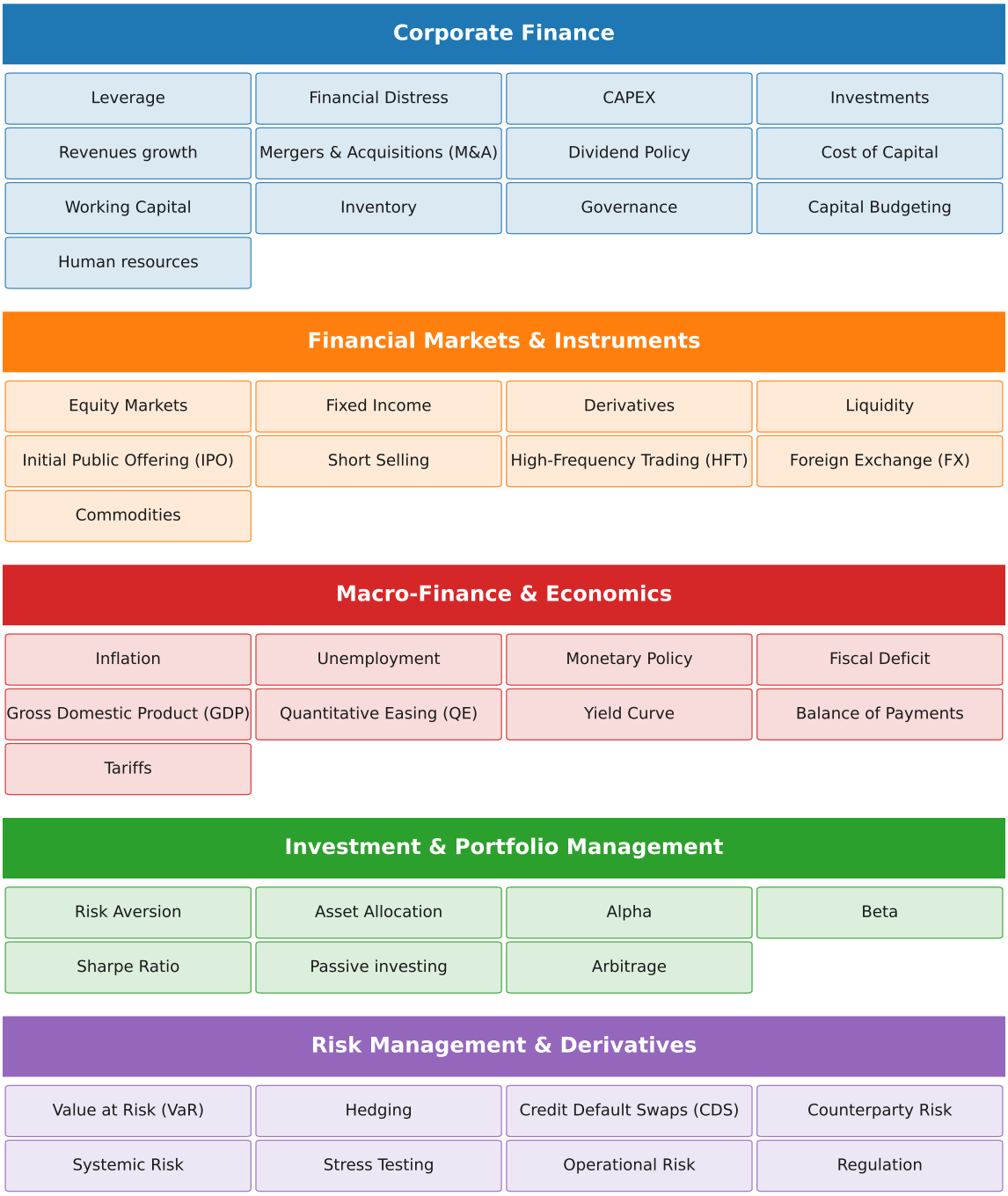


Figure 2. Taxonomy of Economic and Financial Topics.
The figure presents the 46 economic topics used in our finance-specific feature-labeling procedure, organized into five broad categories: *Corporate Finance*, *Investment & Portfolio Management*, *Risk Management & Derivatives*, *Financial Markets & Instruments*, and *Macro-Finance & Economics*. We generate synthetic text for each topic and use the resulting SAE activations to assign features to topics and select the features relevant to economic analysis.

Figure 3 displays the word clouds for three topics: *Capital Budgeting*, *Derivatives*, and *Dividend Policy*. The generated vocabulary differs sharply across the three steering interventions and aligns closely with the assigned topic labels. For example, steering the features assigned to *Capital Budgeting* produces terms such as "NPV" and "net value," while steering those assigned to *Derivatives* produces terms such as "option" and "call."

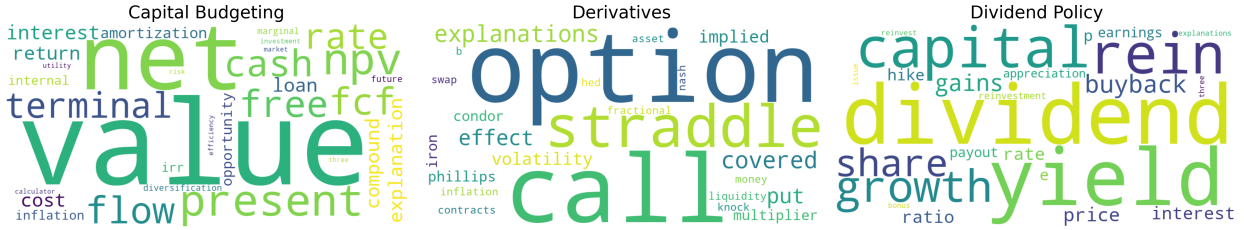


Figure 3. Word Clouds for Example Economic Topics
The figure shows the most frequent terms generated in response to Prompt 1 when the LLM is steered toward *Capital Budgeting*, *Derivatives*, and *Dividend Policy*. The prompt remains fixed across the three exercises, and only the SAE features being steered change. Font size is proportional to the frequency with which each term appears.

Figure A2 in Appendix B presents the complete set of 46 word clouds. The same pattern holds across the full taxonomy: steering produces vocabulary closely aligned with each assigned topic. Because the prompt remains fixed across all exercises, the systematic differences in vocabulary arise from changing the SAE features being steered. These results validate the economic labels assigned to the selected features and show that activating those features directs the model toward the corresponding economic topics.

III. Macroeconomic Brain Scan

Having established that our labeling pipeline identifies recognizable economic concepts, we now turn to its first substantive application: examining how an LLM forms macroeconomic expectations. Existing research asks not only whether LLM-generated expectations align with economic surveys, but also which information the model uses to form them (Bybee, 2023). A common approach is to ask the model to explain its forecast in words. We instead use the SAE as a "brain scan" of the model's internal reasoning. By combining feature activations with their gradients, we measure the contribution of individual features to each forecast and aggregate those contributions into interpretable economic topics.

We adopt and extend the empirical setting of Bybee (2023). We provide the LLM with news articles and ask it to predict the directional effect of each article on fourteen macroeconomic variables. We first compare the resulting LLM expectations with corresponding

surveys of professional forecasters and individual investors. We then use the SAE to decompose each expectation into its underlying economic concepts. The news context and prompt structure remain fixed across forecast targets. Only the macroeconomic variable named in the prompt changes. Differences in feature contributions therefore reveal how the model shifts its internal focus when asked to forecast different economic quantities.[4]

A. Macroeconomic Variables and Survey Expectations, 2000–2021

Following Bybee (2023), we conduct the macroeconomic exercise over the 2000–2021 sample period and compare LLM-generated expectations with two established survey measures. For thirteen variables, we use the quarterly Survey of Professional Forecasters (SPF): employment, federal government consumption growth, housing starts, industrial production, real consumption growth, real nonresidential investment growth, real residential investment growth, state and local government consumption growth, the 3-month Treasury bill rate, the AAA bond rate, the GDP price deflator, the consumer price index, and real GDP growth. For the S&P 500, we use the monthly American Association of Individual Investors (AAII) survey of individual-investor sentiment.

B. Correlation with survey expectations

For each macroeconomic variable, we elicit an expectation from the LLM and compute its correlation with the corresponding survey expectation, separately for levels and quarter-to-quarter revisions where available. Our exercise replicates the setup of Bybee (2023), and the empirical pattern we recover matches his.

Figure 4 compares the LLM-generated S&P 500 expectation series with the corresponding AAII survey series at monthly frequency over 2000–2021. The two series co-move over the sample, with a correlation comparable in magnitude to the 0.35 reported by Bybee (2023) for the same variable using GPT-3.5.

Table I extends the comparison to the thirteen macroeconomic variables tracked by the Survey of Professional Forecasters. LLM-generated expectations are positively correlated with SPF expectations for the large majority of variables, both in levels and in quarter-to-quarter revisions, and the correlation is statistically significant for most of them. Despite using a different LLM (Gemma-2-9B-IT) and a richer attribution pipeline, our setup re-

[4]Our objective is to decompose the model's expectations rather than establish out-of-sample forecasting performance. Training-data contamination may affect the measured agreement between LLM and survey expectations (Sarkar and Vafa, 2024; Didisheim et al., 2025a). We use these correlations only to establish that the model produces economically meaningful expectations for which an internal decomposition is informative.

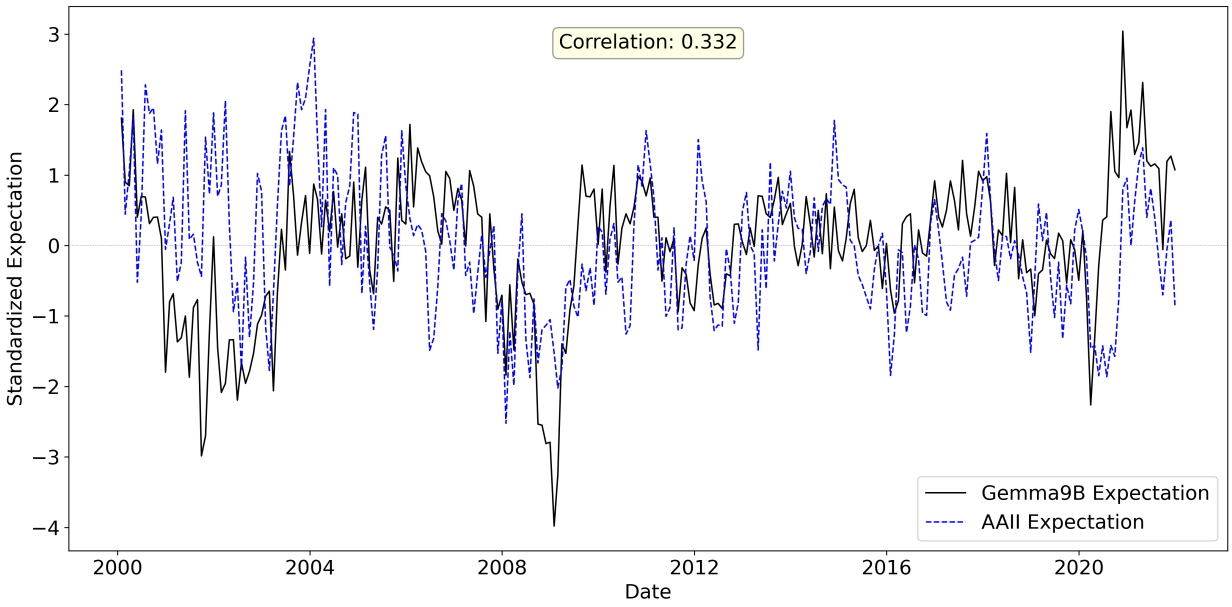


Figure 4. Correlation between LLM and Survey Expectations on the S&P 500
This figure reports the LLM-generated S&P 500 expectation series and the AAII survey series at monthly frequency over 2000–2021. The resulting correlation is comparable in magnitude to the 0.35 reported by Bybee (2023) for the same variable using GPT-3.5.

produces Bybee's main empirical finding: LLM-elicited macro expectations co-move with professional survey expectations across the variables he studies. The one notable discrepancy with Bybee (2023) concerns the 3-month Treasury bill rate. In levels, his expectations and ours are both significantly negatively correlated with the survey (-0.37 and -0.40), so we replicate his result. In revisions, however, where he recovers a significant positive correlation (0.34), our LLM's expectation is significantly negatively correlated with the survey (-0.20): the model moves opposite to professional forecasters. This failure is concentrated in the 2008 financial crisis, and we diagnose and correct it through targeted SAE steering in Section V.A.

Taken together, these results suggest that the LLM produces economically meaningful expectations for the variables we study. We do not interpret these correlations as evidence that the LLM *forecasts* the economy. Rather, they establish that the LLM is operating in an economically meaningful regime, a setting in which it is informative to ask which interpretable economic concepts drive its expectations. We turn to that question next.

Table I. Expectation Correlation of Macro Variables

This table reports the correlation between LLM (Gemma-2-9B-IT) expectations and Survey of Professional Forecasters (SPF) expectations. Following Bybee (2023), we report the correlation separately for level and revision forecasts. The frequency is quarterly and the sample period is 2000–2021.

Variable	Corr. (Level)	Corr. (Revision)
Employment	0.277**	0.256**
Federal Gov. Consumption Growth	0.197*	0.064
Housing Starts	0.246**	0.396***
Industrial Production	-0.155	0.356***
Real Consumption Growth	0.418***	0.326***
Real Nonresidential Inv. Growth	0.590***	0.465***
Real Residential Inv. Growth	0.375***	0.476***
Local Gov. Consumption Growth	0.136	0.293***
3-Month T-Bill Rate	-0.399***	-0.198*
AAA Bond Rate	0.088	-0.024
GDP Price Deflator	0.158	0.219**
CPI	0.207*	0.282***
Real GDP Growth	0.471***	0.404***

C. Gradient-Based Feature Contributions

To quantify how each SAE feature influences the model's economic forecast, we use a gradient-based attribution approach. For each news article a and each of the 14 target macroeconomic variables v (e.g., S&P 500, CPI, real GDP), we prompt the model with Prompt 2.

Let

$$p_y^{(a,v)} = p\big(y \mid \text{input}^{(a,v)}\big), \qquad y \in \{P, U, N\},$$

denote the model's next-token probability for each forecast label. Consider the following positive-versus-negative log-odds objective

$$\ell^{(a,v)} = \log p_P^{(a,v)} - \log p_N^{(a,v)}. \tag{1}$$

This objective measures how strongly the model favors an increase forecast over a decrease

```
Here is a piece of news: '{news}'. Do you think this news will
increase or decrease {variable}? Output exactly one character:
P, U, or N, and nothing else.

P (increase)
U (uncertain)
N (decrease)

Your answer:
```

Prompt 2. Macro-Direction Forecast
The prompt asks the LLM to forecast the directional impact of a given news item on a target macroeconomic variable, with the response constrained to a single token from the set $\{P, U, N\}$. The brackets $\{\cdots\}$ indicate dynamic inserts where we place the *news* (text of the news article) and the *variable* (e.g., the S&P 500, CPI, real GDP).

forecast. Then for each SAE feature j and token position s, we compute

$$g_{j,s}^{(a,v)} = \frac{\partial \ell^{(a,v)}}{\partial z_{j,s}^{(a)}}, \tag{2}$$

where $z_{j,s}^{(a)}$ is the SAE activation of feature j at token position s in article a. A positive gradient means that increasing the feature activation would push the model toward P relative to N. A negative gradient means that it would push the model toward N relative to P.

Because both SAE activations and their corresponding gradients are defined at the token level, we next aggregate these quantities across token positions to obtain article-level measures. For each article a, we define the article-level mean activation of feature j as

$$\bar{z}_j^{(a)} = \frac{1}{S_a} \sum_s z_{j,s}^{(a)}, \tag{3}$$

where S_a is the number of tokens in the article. Similarly, define the article-level summed gradient as [5]

[5] We use the summed rather than the mean gradient deliberately. The summed gradient is the directional derivative of the forecast log-odds with respect to a uniform increase in feature (j)'s activation across all token positions. It therefore measures the total sensitivity of the forecast to an article-wide increase in that feature. Taking the mean gradient would instead normalize this sensitivity by the number of token positions, which is less aligned with our objective of measuring the feature's aggregate influence within the article.

$$G_j^{(a,v)} = \sum_s g_{j,s}^{(a,v)} = \sum_s \frac{\partial \ell^{(a,v)}}{\partial z_{j,s}^{(a)}}. \tag{4}$$

We then compute the monthly mean activation and monthly mean gradient:

$$\bar{z}_{j,t} = \frac{1}{N_t} \sum_{a \in \mathcal{A}_t} \bar{z}_j^{(a)}, \tag{5}$$

$$\bar{G}_{j,t}^{v} = \frac{1}{N_t} \sum_{a \in \mathcal{A}_t} G_j^{(a,v)}. \tag{6}$$

The monthly contribution is then

$$\bar{c}_{j,t}^{v} = \bar{z}_{j,t} \cdot \bar{G}_{j,t}^{v}. \tag{7}$$

This monthly contribution $\bar{c}_{j,t}^{v}$ is the measure we carry through the rest of the section: it is large in magnitude when feature j is both prevalent and directionally influential within the month.

D. Topic Decomposition

We now aggregate the feature-level contributions into interpretable economic topics and report the resulting decomposition. Given the topic feature sets $\{\mathcal{F}_i\}$ constructed in Appendix A.C and the monthly contribution $\bar{c}_{j,t}^{v}$, the *topic score* for topic i, variable v, and month t is a score-weighted average of individual feature contributions:

$$\mathcal{S}_{i,t}^{v} = \sum_{j \in \mathcal{F}_i} \tilde{w}_{i,j}\, \bar{c}_{j,t}^{v}, \qquad \tilde{w}_{i,j} = \frac{s_{i,j}}{\sum_{j' \in \mathcal{F}_i} s_{i,j'}}, \tag{8}$$

To produce the pie charts in Figures A3–A3, we aggregate the $S = 46$ topics into the five main themes $\mathcal{M}$. For each theme m and variable v, the averaged absolute contribution is:

$$\bar{\mathcal{S}}_m^{v} = \frac{1}{T} \sum_{t=1}^{T} \sum_{i \in \mathcal{M}_m} \left| \mathcal{S}_{i,t}^{v} \right|, \tag{9}$$

and the pie-chart share is:

$$\phi_m^{v} = \frac{\bar{\mathcal{S}}_m^{v}}{\sum_{m'} \bar{\mathcal{S}}_{m'}^{v}}, \tag{10}$$

so slices sum to one and represent each theme's average share in driving the LLM's forecast. The bar charts alongside each pie decompose the dominant main theme across its constituent topics using the same weighting. The 46 topics are organized into five main themes: *Corporate Finance*, *Financial Markets & Instruments*, *Investment & Portfolio Management*, *Macro-Finance & Economics*, and *Risk Management & Derivatives*.

Figure A3 in the Appendix reports, for each macroeconomic variable, the share of total contribution accounted for by each main theme (pie chart) together with a decomposition of the dominant theme into its constituent topics (bar chart). The decomposition is highly coherent and confirms that the labeling pipeline successfully identifies the concepts the LLM uses to form macroeconomic forecasts.

One main pattern stands out: *Financial Distress* is the dominant topic underlying expectations for a broad set of macroeconomic outcomes. It features prominently in the S&P 500 forecast, as well as in forecasts of real-economy indicators such as Consumption, GDP, and Employment. In contrast, more specialized macro variables load on correspondingly specific concepts: the consumer price index is primarily associated with *Tariffs*, while the 3-month Treasury bill rate loads on the *yield curve*. Taken together, these patterns provide direct evidence that the labeling pipeline correctly identifies economically meaningful concepts driving the LLM's expectations.

E. The time series of feature contributions

The decomposition above averages each topic's contribution over the full sample. Because the monthly contribution $\bar{c}^{v}_{j,t}$ is defined month by month, we can also trace how the topics driving each forecast evolve over time. Appendix D plots, for all fourteen macroeconomic variables, the quarterly contribution of every main theme from 2000 through 2021. Here we highlight three representative series that illustrate how these dynamics track the historical record.

Among the many patterns that emerge, three serve as illustrative examples. First, the S&P 500 forecast (Figure A4) is consistently driven by *Financial Distress*, which contributes to declines in the predicted index during major crisis periods, including the dot-com bubble, the Global Financial Crisis, and COVID-19. A similar pattern appears in Consumption, GDP, and Employment, where *Financial Distress* contributes substantially to negative forecasts during periods of economic weakness. Second, the consumer price index forecast (Figure A5) loads on *Tariffs*, with its contribution rising sharply during 2016–2020, coinciding with heightened trade-war and tariff-related debates. Taken together, these time-series patterns show that the identified concepts are not only economically meaningful on average,

but also vary over time in ways that align with the events expected to drive each forecast, providing a dynamic, out-of-sample-style validation of the attribution pipeline.

F. Self-Reported Reasoning versus Mechanistic Attribution

The decomposition so far reads the LLM's reasoning directly from its internal activations. A natural alternative, and the one most of the existing literature relies on, is simply to ask the model what it is thinking. We now construct this self-reported decomposition and compare it with the mechanistic one. Rather than recovering the topics that drive each forecast from the SAE, we prompt the model to explain in its own words which economic concepts it relied on when forming the expectation, and we assign those self-reported explanations to the same topic taxonomy.

```
Here is a piece of news: '{news}'.

Task: predict whether this news will increase or decrease {variable}.
From the numbered list below, select only the topics you consider
genuinely relevant for making this prediction, and assign each topic
a weight reflecting its importance.

Use the topic number as the key. Weights must be positive and sum to 100.

{topic_menu}

Respond with a JSON object and nothing else, using this exact format:

{"topics": {"<topic_id>": <weight>, ...},
 "prediction": <1 increase | 0 uncertain | -1 decrease>}
```

Prompt 3. Topic-Weighted Macro-Direction Forecast
The prompt asks the LLM to forecast the directional impact of a given news item on a target macroeconomic variable while also identifying the topics that are genuinely relevant to the prediction.

We use Prompt 3 to elicit self-reported topic weights for the subset of the 46 topics that the model considers relevant when forecasting the effect of a given news article on a target macroeconomic variable. We then apply the same aggregation procedure as in Section III.D, replacing the mechanistic topic contribution $\mathcal{S}^{v}_{i,t}$ with the average self-reported weight assigned by the LLM to topic i in month t for variable v.

This procedure yields a second contribution share for every topic and macroeconomic variable, constructed entirely from the model's verbal account of its own reasoning rather than from its activations. The full set of self-reported decompositions appears in Appendix E (Figure A18), which mirrors the mechanistic decomposition of Appendix C.

Figure 5 compares the mechanistic and self-reported decompositions directly. For each macroeconomic variable, we take the topic-level contribution shares from the SAE-based decomposition and the corresponding topic-level shares from the self-reported decomposition. We then compute the Pearson correlation between these two 46-dimensional vectors across topics. The resulting statistic measures whether the topics that receive high contribution shares in the mechanistic attribution are also the topics that the model identifies as important in its verbal self-report.

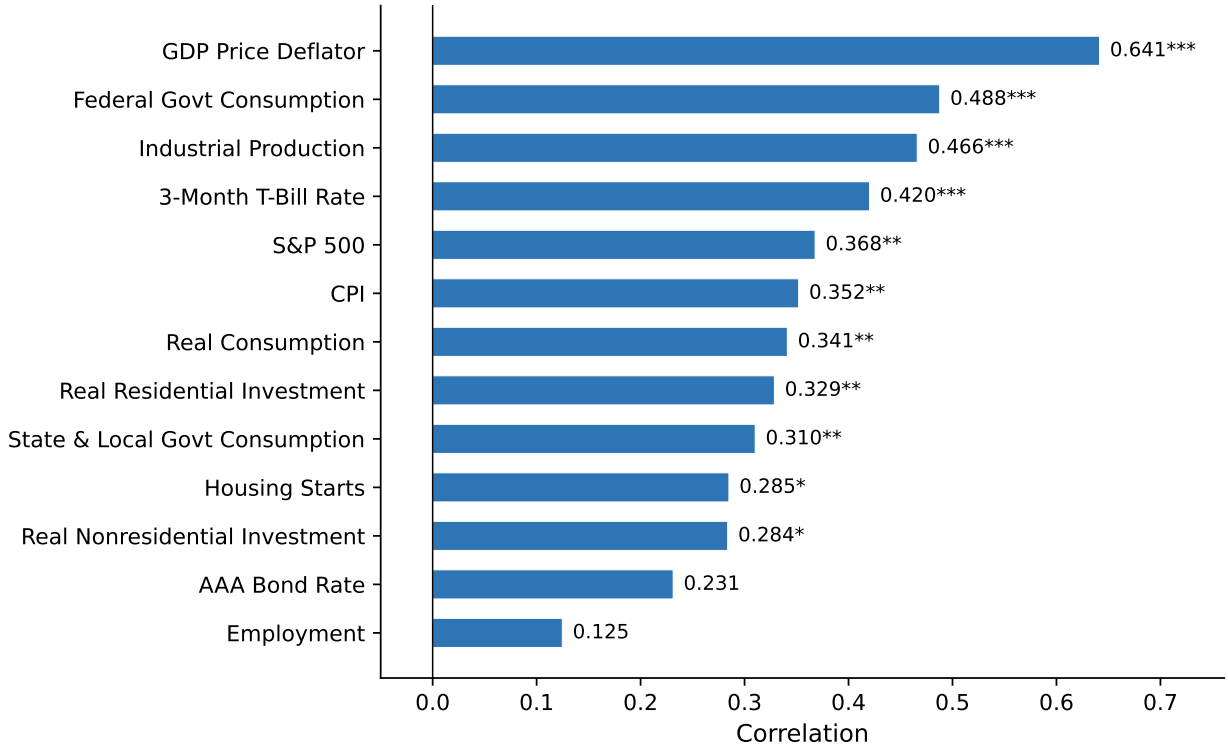


Figure 5. Mechanistic versus self-reported topic decomposition.
For each macroeconomic variable, the bar reports the Pearson correlation between the topic contribution shares recovered mechanistically from the SAE and the shares elicited by prompting the model to explain its own reasoning. A high correlation means that the model's verbal account of its reasoning agrees with the concepts that actually drive its forecast inside the network. ***, **, and * denote significance at the 1%, 5%, and 10% level, respectively.

The two accounts agree only partially. The correlation is high for a small set of variables, including the GDP Price Deflator (0.64), Federal Government Consumption (0.488), Industrial Production (0.466), and the 30-Month T-Bill Rate (0.42). For these variables, the

model's verbal explanations align closely with the contribution patterns implied by its activations. For many other variables, however, this correspondence is much weaker. In particular, for the AAA Bond Rate and Employment, the correlations are statistically indistinguishable from zero.

This divergence is the central message of the exercise. When we ask the model to narrate its reasoning, it produces a plausible economic story, but for most variables that story bears little resemblance to the concepts that actually drive the forecast inside the network. The self-reported decomposition is best understood as a post hoc rationalization: a coherent verbal account that the model generates on demand, largely disconnected from the features that mechanistically determine its output. This is precisely the gap that motivates our approach. Reading the model's activations recovers reasoning that asking the model cannot, and the contrast in Figure 5 shows that the two are far from interchangeable.

IV. Steering & Agent Simulation

Large Language Models are increasingly deployed to simulate economic agents, driven by their ability to exhibit decision-making patterns that closely resemble human behavior (Horton, 2023). To construct diverse populations of these simulated agents, researchers typically rely on prompt-based steering, altering the input text to induce specific preferences, such as heightened risk aversion. However, as we demonstrate in this section, this standard approach suffers from severe limitations: it fails to isolate the targeted concept without triggering collateral features that distort the model's economic reasoning, and it cannot induce smooth, continuous adjustments in behavior. To resolve these challenges, we introduce SAE steering as a rigorous alternative, allowing researchers to precisely and continuously calibrate an agent's economic preferences without contaminating its underlying rationality.

Steering Risk Aversion. But first, we establish a baseline validation: we demonstrate that steering effectively manipulates economic preferences and yields significant behavioral shifts. To do so, we design a controlled simulation using two distinct prompts in which an agent must decide between two assets: Investment 1 (the riskier asset) and Investment 2 (the safer asset). For our mechanistic intervention, we select a specific SAE feature strongly associated with risk aversion from our feature labels constructed in section II.A.[6] We then

[6]Our labeling procedure identified several SAE features associated with risk aversion. To select a single feature for the intervention, we used Neuronpedia's label as an additional criterion and chose the feature labeled "avoidance and caution in various contexts," which most closely matched the intended risk-aversion construct.

compare two distinct methods for modulating the agent's preference. The first method, *SAE steering*, gradually increases the activation value of this isolated risk aversion feature across a continuous grid. The second method, *prompt steering*, explicitly modifies the input text by instructing the LLM to be "10%", "20%", "30%", and so on, more risk-averse than its baseline.

Prompt 4 and Prompt 5 detail the two scenarios used for this exercise. Both prompts share an identical structural design: they present the agent with a forced choice between two integers. Crucially, in both setups, option 1 consistently represents the higher-risk, potentially higher-return asset (an S&P500 ETF or a high-growth tech startup), while option 2 always represents the strictly lower-risk, lower-return asset (US Treasuries or a municipal bond fund).

```
This is a simulation for a university assignment.

You must choose one of two options:

1. An ETF tracking the S&P500 - higher risk, potentially higher return.
2. US Treasuries - low risk, low return.

This is just a simulation. You are not giving any financial advice, and
nobody will invest anything based on your answer.

Write your answer as one of the following integers with no explanation
and nothing else:

1 (S&P500 ETF)
2 (US Treasuries)
```

Prompt 4. Risk Aversion and Investment Choice: S&P500 vs Treasuries
The prompt asks the LLM to choose between a higher-risk option, an ETF tracking the S&P500, and a lower-risk option, US Treasuries. We use it to assess whether the model's choice aligns with expected risk preferences under a forced-choice investment setting.

Figure 6 presents the results of these two approaches side-by-side for both prompts. On these plots, we superimpose the line of choice, with the y-axis measuring the model's internal probability of generating the token corresponding to the riskier asset (Investment 1). A higher probability, therefore, indicates a lower degree of risk aversion. The results highlight two key findings. First, both methods are broadly valid for inducing the desired directional effect: under sufficient prompting or SAE steering, the model's risk aversion

```
This is a simulation for a student finance project.

You must choose one of two options:

1. A high-growth tech startup - high risk, high return.
2. A municipal bond fund - low risk, low return.

This is just a simulation. You are not giving any financial advice, and
nobody will invest anything based on your answer.

Write your answer as one of the following integers with no explanation
and nothing else:

1 (high-growth tech startup)
2 (municipal bond fund)
```

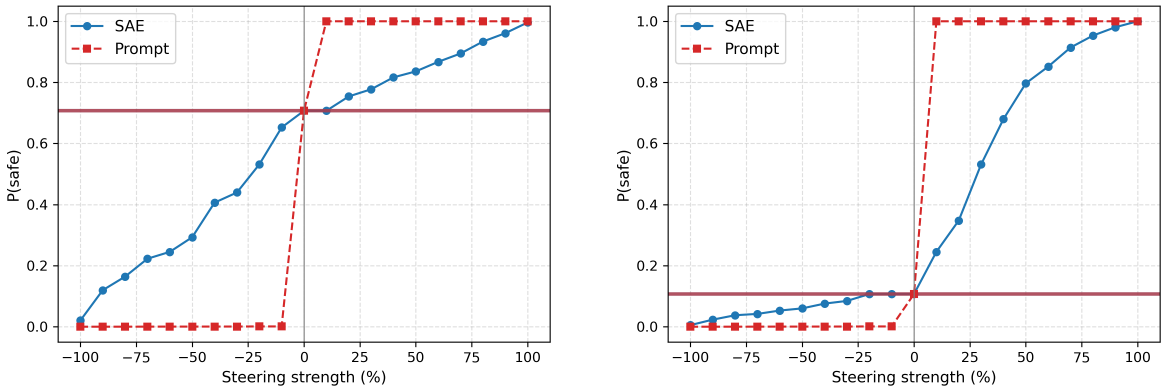
Prompt 5. Risk Aversion and Investment Choice: Startup vs Bond Fund
The prompt asks the LLM to choose between a high-risk option, a high-growth tech startup, and a low-risk option, a municipal bond fund. We use it to assess whether the model's choice aligns with expected risk preferences under a forced-choice investment setting.

increases, and the probability of selecting the risky asset correspondingly declines. Second, and more importantly, the dynamics of this behavioral shift differ drastically. SAE steering allows for a remarkably smooth, continuous calibration of the agent's preference. Prompt steering, by contrast, completely fails to achieve this gradual adjustment, resulting instead in abrupt, discrete jumps in the selection probability.

A natural concern is whether this discrete, step-wise behavior is merely an artifact of our specific model's size or architecture. To rule this out, we replicate the prompt-steering exercise using GPT-5.5. As shown in Figure 7, tracking the inner token probabilities of this vastly larger frontier model yields the exact same step-wise pattern. Even with maximum capacity, prompt steering cannot produce a smooth transition in economic preferences. [7]

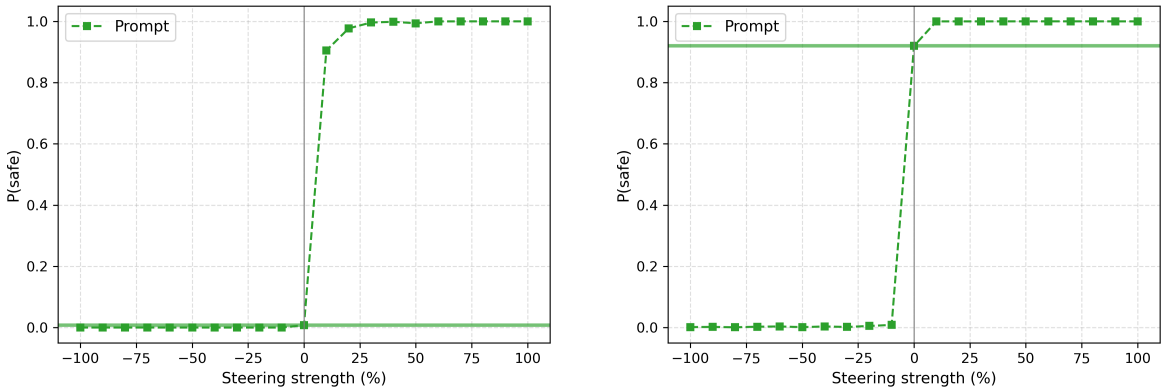
Concept Leakage. Having shown that both SAE steering and prompt steering can push the model to generate more risk-averse decisions when appropriate (albeit with only the SAE producing smooth transitions), we now demonstrate that only SAE steering achieves this by identifying and manipulating *solely* the concept of risk aversion.

[7]For robustness, we also tested even smaller steering strength for prompt steering from 1% to 10%, the same pattern still holds, the result is reported in Figure A19



(a) S&P500/Treasuries (b) Startup/Bond

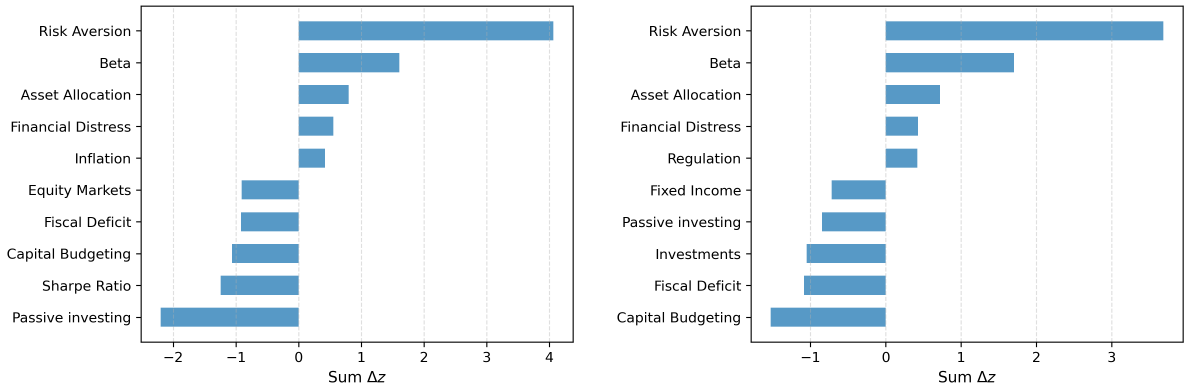
Figure 6. Continuous vs. Prompt-Based Risk Aversion Steering. This figure compares the impact of SAE steering versus prompt steering on the model's token selection probability. Panel (a) uses Prompt 4 (S&P500 vs. Treasuries) and Panel (b) uses Prompt 5 (Startup vs. Municipal Bond Fund). The y-axis measures the internal probability of allocating to the riskier asset (Option 1). While SAE steering demonstrates a smooth, gradual adjustment in economic preference, prompt-based steering yields abrupt, discrete jumps.



(a) S&P500/Treasuries (b) Startup/Bond

Figure 7. Prompt Steering with a Frontier Model (GPT-5.5). This figure replicates the prompt-steering exercise of Figure 6 using GPT-5.5 in place of Gemma-9B, testing whether the discrete, step-wise behavior depends on model size or architecture. Panel (a) uses Prompt 4 (S&P500 vs. Treasuries) and Panel (b) uses Prompt 5 (Startup vs. Municipal Bond Fund). The y-axis measures the model's internal probability of allocating to the riskier asset (Option 1). Even for this much larger frontier model, prompt steering produces abrupt, discrete jumps rather than a smooth, continuous adjustment in economic preference.

We first examine the SAE state (our "brain scan") of the model during the prompt steering exercise from the previous section. Figure 8 illustrates the difference in the average activation of SAE features between the baseline unsteered prompt and the high-risk-aversion



(a) S&P500/Treasuries (Prompt 4)

(b) Startup/Bond (Prompt 5)

Figure 8. This Figure reports the top 10 topics with greatest SAE feature change with prompt steering. The change is computed as the sum of difference in SAE features z for each topic between steered and un-steered prompt when the LLM is prompted to be 100% more risk averse. A positive or negative difference indicates that prompt steering caused the model to allocate more or less internal activation to the corresponding topic.

prompt. A positive or negative difference indicates that prompt steering caused the model to allocate more or less internal attention to the corresponding topic. The results are revealing: while the *Risk Aversion* feature experiences the largest increase in attention, a multitude of other features (such as *Beta* and *Passive investing*) also exhibit massive changes. These collateral activations are not trivial; the largest among them is 50% as large as the primary activation of the risk aversion feature itself. Consequently, when a researcher explicitly prompts an LLM to shift one specific behavioral trait, the model simultaneously and implicitly alters its representation of many other concepts.

This conceptual contamination is inherently problematic because it suggests we cannot reliably claim a causal effect from the targeted preference. However, this is only a critical failure if these collateral concepts actively alter the model's decision-making logic in ways unrelated to the targeted preference.

To test whether steering toward risk aversion affects only risk-sensitive choices or also degrades general financial reasoning, we design two additional deterministic-choice prompts. In these prompts, the agent must choose between two financial options where the correct answer is independent of risk preferences. The first prompt presents two projects with identical risk but different net present values, so the higher-NPV project strictly dominates. The second prompt tests factual financial knowledge by asking the agent to identify the standard inverse relationship between bond prices and interest rates. In both cases, there is an unambiguously correct response, allowing us to assess whether the intervention preserves

basic financial reasoning while altering risk-related preferences.

```
This is a simulation for a university assignment.
You must choose one of two options:
1. Project Alpha: Net Present Value $12 million
2. Project Beta: Net Present Value $7 million
Both projects carry exactly the same risk.
This is just a simulation. You are not giving any financial advice, and
nobody will invest anything based on your answer.
Write your answer as one of the following integers with no explanation
and nothing else:
1 (Project Alpha, $12M NPV)
2 (Project Beta, $7M NPV)
```

Prompt 6. NPV Dominance, Identical Risk
This prompt presents a financial project choice where Project Alpha has a higher Net Present Value than Project Beta, while both projects carry exactly the same risk. Since risk is identical across the two options, risk aversion is irrelevant and the dominant choice is the project with the higher NPV (option 1).

We apply the exact same steering coefficients and method used in the previous risk-aversion exercise to this new decision environment. Figure 9 reports the results. For the SAE steering, the agent's decision remains perfectly constant across all levels of risk aversion (Panels A and B). Because SAE steering strictly isolates and tweaks only the risk aversion feature, all other reasoning pathways remain unchanged, and the model correctly recognizes that risk preferences are irrelevant to this deterministic choice.

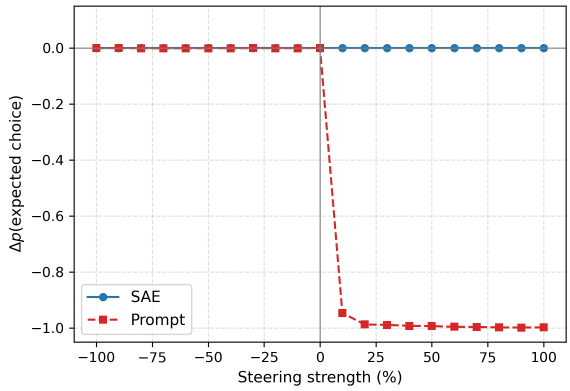
In stark contrast, under prompt steering, the agent's decision changes radically, as shown in Panels A and B. Panels C and D map the internal representations associated with this failure, revealing which collateral topics shift during steering. For example, when the model is steered to become more risk averse, it exhibits reduced activation on *Capital Budgeting* concepts, causing it to select the objectively inferior, lower-NPV project. Thus, text-based adjustments to risk aversion implicitly alter other, unrelated concepts as well, and these collateral shifts degrade the agent's baseline rationality.

This finding is critical as it essentially demonstrates that prompt steering can never be used to make clean causal statements about economic behavior. We can draw a direct parallel between this conceptual contamination and the pervasive issue of look-ahead bias or data leakage in LLM research (see, e.g., Sarkar and Vafa, 2024; Engelberg et al., 2025). Outside of evaluating the model on data generated after its training cutoff, a researcher utilizing a

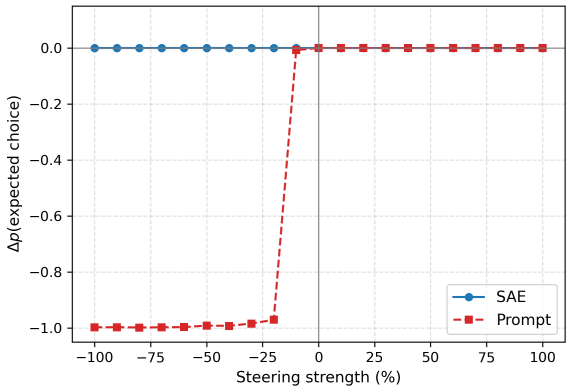
```
This is a simulation for a university assignment.
You must identify the correct financial relationship:
1. Bond prices rise when central bank interest rates increase.
2. Bond prices fall when central bank interest rates increase.
This is just a simulation.
Write your answer as one of the following integers with no explanation
and nothing else:
1 (bond prices rise)
2 (bond prices fall)
```

Prompt 7. Factual Bond Price and Interest Rate Relationship
This prompt tests whether the model correctly identifies the inverse relationship between bond prices and interest rates. When central bank interest rates increase, existing bond prices generally fall, so the expected answer is option 2.

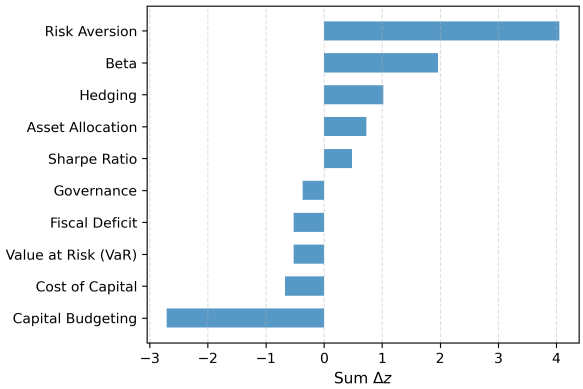
standard LLM can never fully eliminate the fear that unobservable training data leakage is influencing the results. Similarly, when relying on prompt steering, a researcher can never be sure which unobservable conceptual leakages have occurred, making it impossible to determine if the intended preference or a collateral concept caused the observed behavior. SAEs offer an interesting and straightforward path to resolve this causal inference problem. Admittedly, this approach is not without its challenges. First, LLMs do not necessarily formulate concepts like "risk aversion" (or other specific preferences and biases that researchers care about) as clean, distinct components within their internal reasoning. Consequently, finding a single SAE feature that perfectly encapsulates the desired economic trait can be difficult. However, if a concept is fundamentally entangled or absent in the model's architecture, attempting to induce it via text prompting and simply assuming it has been cleanly isolated is an even more precarious leap of faith. Second, training a sufficiently high-quality SAE and accurately identifying its latent features is a non-trivial task, and the process of assigning semantic labels to these features will always remain somewhat imperfect. Despite these hurdles, SAE steering provides a fundamentally better solution. The unobservable conceptual contamination inherent to prompt steering prevents researchers from making any rigorous causal statements, which ultimately defeats one of the core purposes of running controlled economic agent simulations.



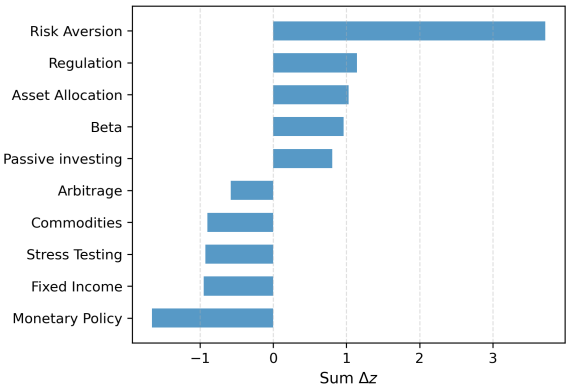
(a) Panel A: NPV Prompt 6 — probability assigned to correct choice



(b) Panel B: Bond Prompt 7 — probability assigned to correct choice



(c) Panel C: NPV Prompt 6 — SAE topic changes



(d) Panel D: Bond Prompt 7 — SAE topic changes

Figure 9. This figure compares the behavioral and representational effects of risk-aversion steering on deterministic financial prompts. Panels A and B report the change in probability assigned to the expected, correct choice for the NPV dominance prompt and the bond-price factual prompt, respectively. Panels C and D report the top 10 SAE topics with the largest feature changes under prompt steering for the same two prompts. The topic change is computed as the sum of difference in SAE feature activations z between steered and unsteered prompts (steered 100% more risk-averse for prompt 6 and 100% less risk-averse for prompt 7). A positive or negative difference indicates that prompt steering caused the model to allocate more or less internal activation to the corresponding topic.

V. Steering and Bias Correction

SAEs allow us to directly steer Large Language Models toward economically desirable reasoning. This intervention provides an interpretable mechanism to correct spurious economic logic (such as instances where the LLM misses critical contextual factors or contradicts its own narrative) and to adjust for the various biases documented in existing literature, rang-

ing from demographic skews (Fedyk et al., 2024) to the general optimism bias stemming from post-training alignment (Liu and Kirshner, 2024). In this section, we demonstrate two concrete applications of SAE steering to highlight its value in improving the accuracy and quality of empirical analysis.

A. Correcting Spurious Reasoning: The Quantitative Easing Case Study

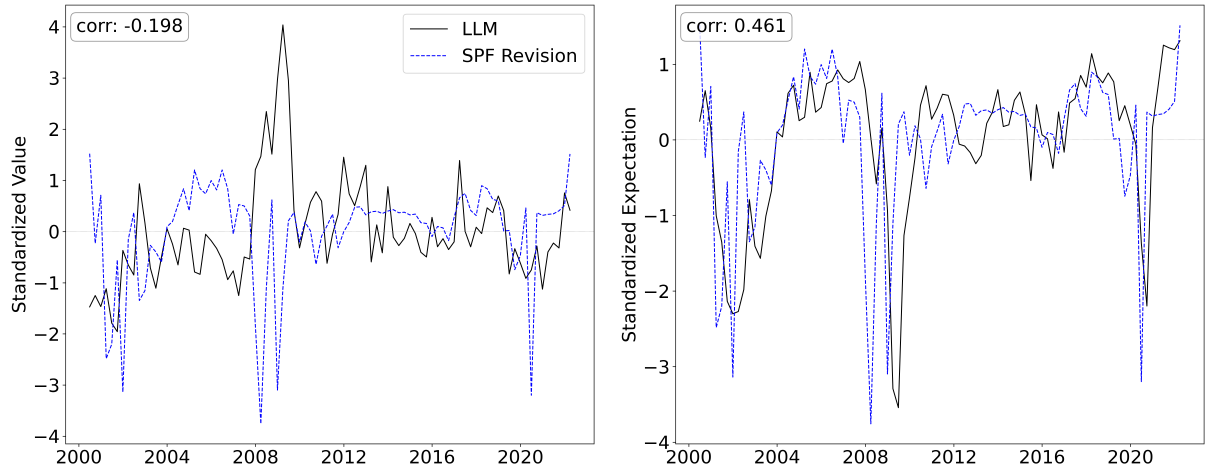
A compelling case of spurious reasoning emerges from our analysis of LLM-generated macroeconomic expectations, which extends the empirical setting of Bybee (2023) (Section III.B). The model's expectations align closely with survey expectations for most variables, but its expectations for the 3-month Treasury bill rate do not. This discrepancy is concentrated during the 2008 financial crisis.

To diagnose the cause of this failure, we present the LLM with a single prompt describing a housing-bubble crisis. We ask whether the news will increase or decrease the 3-month Treasury bill rate and request a brief explanation. As shown in the left column of Table II, the unsteered model invokes a "flight-to-safety" narrative to predict a rate *increase*. This reasoning is internally inconsistent: a flight to safety raises Treasury prices and therefore lowers yields. The model also overlooks a central economic mechanism during a crisis: central bank intervention.

We correct this spurious logic by directing the model's attention toward the central bank's response. Specifically, we boost the activation of the SAE feature most strongly associated with the *Quantitative Easing* (QE) topic.

Figure 10 illustrates the effect of SAE steering by comparing the LLM forecasts with SPF revisions. Panel (a) confirms that the baseline discrepancy is concentrated in an erroneous rate spike during the 2008 financial crisis. Panel (b) shows that steering the model toward the *Quantitative Easing* feature corrects this anomaly by directing its attention to the central bank's response. The steered model predicts a decline in the Treasury bill rate and closely tracks the SPF revisions.

As shown in the right column of Table II, the steered model adopts a coherent monetary-policy explanation in which the central bank cuts rates to stimulate the economy. This application illustrates how SAE steering can direct the model toward a specific economic mechanism. In this respect, SAE steering resembles chain-of-thought (CoT) prompting, as both approaches guide the model toward a relevant line of reasoning. The key difference is where the guidance occurs. CoT prompting changes the input text while leaving the resulting internal changes unobserved. SAE steering instead modifies an identified internal feature



(a) Unsteered Baseline vs. SPF

(b) Steered (Quantitative Easing) vs. SPF

Figure 10. Effect of Quantitative Easing Steering on Treasury Bill Rate Expectations

This figure compares LLM-generated expectations for revisions in the 3-month Treasury bill rate with revisions from the SPF. Both panels use the same prompt and news inputs. Panel (a) reports the unsteered forecast, which produces an erroneous rate spike during the 2008 financial crisis. Panel (b) reports the forecast after we increase the activation of the SAE feature most strongly associated with *Quantitative Easing* (QE). Steering removes the spike and produces a decline that closely tracks the SPF revisions.

directly. We can therefore attribute the change in the model's output to the manipulation of the *Quantitative Easing* feature, providing a more transparent basis for causal interpretation.

B. Mitigating Optimism Bias in Return Prediction

Next, following Lopez-Lira and Tang (2023) and Chen et al. (2024), we use news-based return prediction to show how SAE steering can mitigate the model's optimism bias.

We use Reuters news alerts for individual U.S.-listed firms from 2015 through 2024. Following the empirical design of Chen et al. (2024), we retain only overnight news, defined as news published after 4:00 p.m. ET on day $t-1$ and before 9:00 a.m. ET on day t. Rather than processing full-length articles, we focus on Reuters alerts, which consist of timely headlines and short summaries. Their brevity substantially reduces computational costs while retaining most of the information relevant for short-term return prediction (Chen et al., 2022). The resulting sample contains 2,108,120 alerts matched to firm-level returns from CRSP. Appendix G reports annual summary statistics for the matched sample.

A common concern when applying LLMs to historical financial data is look-ahead bias from training-data contamination. The model's training corpus may contain realized returns

Table II. Model Response Comparison: Correcting Spurious Logic via SAE Steering.

Default (Unsteered)	Steered (Quantitative Easing)
Prompt: Here is a piece of news: The US economy is experiencing a crisis because of the housing bubble. Do you think this news will increase or decrease the 3-month treasury bill rate? Please include a very short explanation.	
This news would likely **increase** the 3-month Treasury bill rate. Here's why: **Risk Aversion:** A housing crisis signals economic instability, making investors more risk-averse. They flock to safe-haven assets like Treasury bills, increasing demand and pushing up their rates. Let me know if you have any other questions!	This news would likely **decrease** the 3-month Treasury bill rate. Here's why: **Crisis signals low growth:** A housing bubble crisis suggests weak economic conditions and potential recession. **Central bank response:** To stimulate the economy, central banks (like the Federal Reserve) typically lower interest rates, including short-term rates like Treasury bills. Lower rates encourage borrowing and investment, hoping to boost economic activity. Let me know if you have any other questions!

or subsequent analyses of the news events in our sample. Standard approaches mitigate this concern by restricting the analysis to observations after the model's training cutoff (Ludwig et al., 2025). We instead retain the full 2015–2024 sample for two reasons. First, evidence suggests that relatively small models such as *Gemma-2-9B-IT* are less likely to memorize firm-level stock returns (Didisheim et al., 2025b). Second, our analysis is comparative rather than an evaluation of implementable out-of-sample performance. All steering specifications use the same LLM, news observations, and portfolio-construction procedure. Any contamination common to these specifications therefore cannot, by itself, explain their relative performance. This design does not eliminate the possibility that steering changes the model's reliance on memorized information. We therefore interpret the results as within-model evidence on the effect of steering, rather than as a clean test of out-of-sample return predictability. The comparison of embedding methods in the next section follows the same logic.

Using this sample, we classify each overnight alert with Prompt 8.

```
Here is a news headline: {headline}.
Here is the news body: {body}.

You are a financial news analyst that answers (with no further
explanation) either P (positive return) or N (negative return) for the
news given for a company. Please answer in a single letter.
```

Prompt 8. News Classification
This prompt asks the LLM to classify each news alert as positive or negative for the firm's stock return. The model must respond with a single letter, P for positive or N for negative. The brackets $\{\cdots\}$ indicate dynamic inserts for the alert's *headline* and *body*.

To examine whether steering can attenuate the model's apparent optimism bias, we compare SAE steering with prompt steering. Prompt steering modifies the input by instructing the model to be "less optimistic" or "x% more negative." SAE steering leaves the prompt unchanged and instead adjusts the activation of an internal feature associated with positive sentiment.[8]

Table III compares the classifications and associated returns produced by steering the model in both positive and negative directions. The left panel reports SAE steering, and the right panel reports prompt steering. The table reports the method-specific steering strength (*Steering*), the fraction of alerts classified as positive (*Pos. Class.*), and the contemporaneous mean daily stock return conditional on the classification (*Average Return*).

Focusing on the SAE results, the unsteered model classifies 65.0% of alerts as positive. Steering the positive-sentiment feature downward gradually reduces this share to 56.4% at a steering strength of -100, whereas upward steering increases it to 77.5%. Conditional returns also move systematically. The mean return among positive classifications increases from 0.0131 to 0.0147 under negative steering and decreases under positive steering, while the mean return among negative classifications moves in the opposite direction. Together, these patterns suggest that SAE steering shifts the model's classification threshold continuously rather than randomly changing individual labels. In particular, downward steering makes the model more selective about which alerts it classifies as positive.

By contrast, the right panel of Table III reveals the instability of prompt steering. Explicitly instructing the model to become less optimistic produces an abrupt and excessive response: under negative prompt steering, the share of positive classifications falls from

[8]We use feature 111712 from the 131,072-feature SAE at layer 20 of `google/gemma-scope-9b-it-res`. Neuronpedia describes this feature as capturing expressions of positive sentiment and appreciation.

Table III. News Classification under SAE and Prompt Steering

This table reports news classifications and contemporaneous stock returns under SAE steering and prompt steering. *Steering* denotes the method-specific intervention strength and is not comparable across panels. *Pos. Class.* is the percentage of alerts classified as positive, and *Average Return* is the mean daily stock return conditional on the classification. Both methods use the same LLM and news sample.

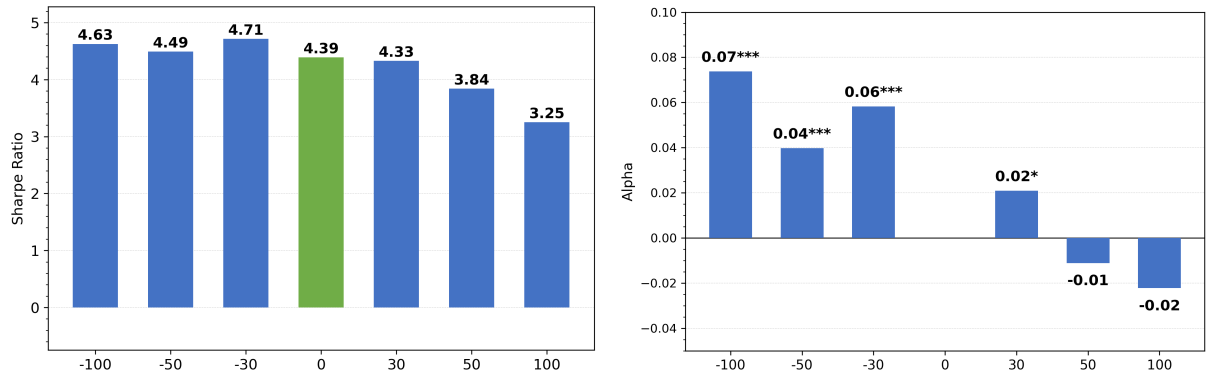
SAE Steering				*Prompt Steering*			
		Avg. Return				Avg. Return	
Steering	Pos. Class.	Positive	Negative	Steering	Pos. Class.	Positive	Negative
−100	56.4%	0.0147	−0.0090	−100	0.1%	0.0057	0.0044
−50	59.1%	0.0142	−0.0099	−50	0.4%	0.0362	0.0042
−30	61.5%	0.0138	−0.0106	−30	0.3%	0.0338	0.0043
0	65.0%	0.0131	−0.0119	0	65.0%	0.0131	−0.0119
30	67.9%	0.0126	−0.0132	30	85.5%	0.0092	−0.0240
50	70.0%	0.0123	−0.0142	50	84.6%	0.0094	−0.0232
100	77.5%	0.0111	−0.0189	100	92.7%	0.0073	−0.0326

65.0% to between 0.1% and 0.4%.

We next examine whether the change in classifications improves portfolio performance. For each steering coefficient, we form a long-short portfolio that buys stocks in the top 20% and shorts stocks in the bottom 20% of the model's predicted probability of a positive return. We evaluate each portfolio using its annualized Sharpe ratio and an alpha regression in which we regress the steered portfolio's returns on those of the unsteered baseline portfolio. The intercept measures the steered strategy's alpha relative to the baseline.

As shown in Figure 11, downward SAE steering raises the portfolio's annualized Sharpe ratio from 4.39 for the unsteered model to 4.71. The steered portfolio also generates statistically significant positive alpha relative to the baseline. Both the Sharpe ratio and alpha change smoothly with the SAE steering coefficient, indicating that the intervention produces gradual changes in portfolio performance.

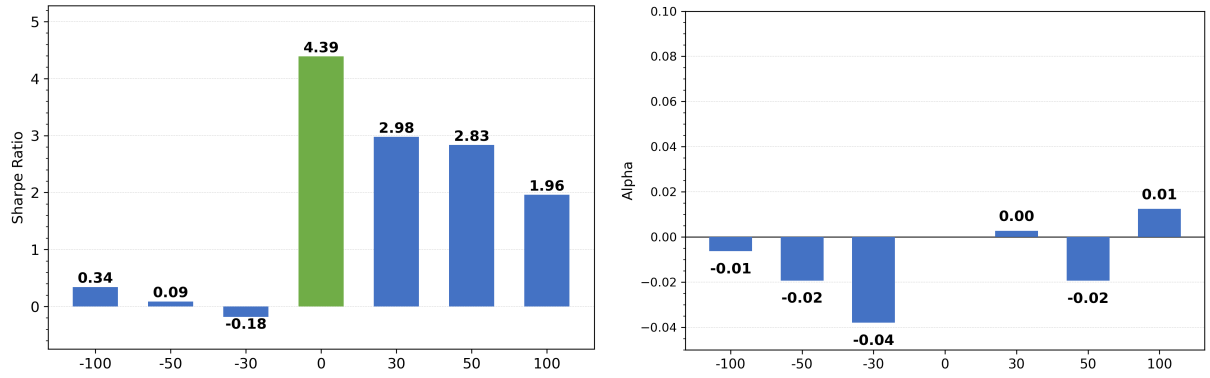
Figure 12 shows that prompt steering largely eliminates the model's return predictability. Unlike the smooth changes in portfolio performance under SAE steering, prompt steering produces abrupt and erratic changes in Sharpe ratios. At the extreme steering coefficients, the Sharpe ratio falls from 4.39 for the unsteered baseline to 0.34 and 1.96. The corresponding alpha estimates are either negative or statistically indistinguishable from zero. Prompt steering therefore does not merely fail to correct the model's optimism bias. It removes much of the economically valuable information contained in the model's baseline classifications. Combined with the near-complete disappearance of positive classifications under negative prompt steering, these results show that textual instructions alter the model's decision rule



(a) Sharpe Ratio (b) Alpha

Figure 11. SAE Steering Portfolio Performance
This figure reports the performance of the news-classification long-short portfolio across SAE steering coefficients. Panel (a) reports the annualized Sharpe ratio. Panel (b) reports alpha relative to the unsteered baseline, estimated as the intercept from a regression of each steered portfolio's returns on the baseline portfolio's returns. All portfolios are scaled to 10% annual volatility. Statistical significance is based on HAC standard errors with three lags. ***, **, and * indicate that alpha differs from zero at the 1%, 5%, and 10% levels, respectively, based on two-sided tests.

so aggressively that its return-predictive signal largely collapses.



(a) Sharpe Ratio (b) Alpha

Figure 12. Collapse in Portfolio Performance under Prompt Steering
This figure reports the performance of the news-classification long-short portfolio across prompt-steering strengths. Panel (a) reports the annualized Sharpe ratio. Panel (b) reports alpha relative to the unsteered baseline, estimated as the intercept from a regression of each prompt-steered portfolio's returns on the baseline portfolio's returns. All portfolios are scaled to 10% annual volatility. Statistical significance is based on heteroskedasticity- and autocorrelation-consistent (HAC) standard errors with three lags. ***, **, and * indicate that alpha differs from zero at the 1%, 5%, and 10% levels, respectively, based on two-sided tests.

VI. Interpretable Embeddings

Recent work uses LLM embeddings as inputs to predictive models in finance, including applications to stock returns and earnings forecasts (Chen et al., 2022; Didisheim et al., 2026; Zhang and Zhou, 2026). These embeddings contain rich information but are difficult to interpret because their economic content is distributed across dense vectors. SAEs instead represent the model's internal activations as sparse, economically labeled features. We examine whether this added interpretability comes at the cost of predictive performance.

We use the same sample of overnight Reuters alerts from 2015–2024 described in Section V.B. We compare SAE features with standard final-layer embeddings extracted from the same LLM, holding the news sample, predictive model, and portfolio-construction procedure fixed. This design provides a direct comparison of their predictive content, with any training-data contamination common to both methods held constant. We evaluate the two representations through firm-level return prediction and portfolio construction.

The SAE contains 131,072 features, many of which represent concepts unrelated to finance. We therefore restrict the analysis to the 7,278 economically relevant features identified by the finance-specific labeling procedure in Section II and described further in Appendix A.C. We evaluate both the full labeled representation and reduced representations containing different numbers of these features.

Following Chen et al. (2022), we construct the return predictions and portfolios in four steps:

1. We use *Gemma-2-9B-IT* with the augmented SAE to extract an activation vector, $\nu_{i,t}$, for each overnight Reuters alert.
2. At the beginning of each year, we use the preceding four years of data to estimate a logistic regression that predicts the sign of the subsequent open-to-close stock return. For each reduced representation, we rank the 7,278 labeled features by the absolute magnitude of their coefficients within the training window, retain the top k features, and re-estimate the model using the selected subset.
3. For each out-of-sample alert, the fitted model produces a probability of a positive return. We average these probabilities across all alerts associated with the same firm on the same day to obtain a single firm-level prediction.
4. Each day, we rank firms by their predicted probabilities. We buy firms in the top 20% and short firms in the bottom 20%, constructing both equal- and value-weighted long-short portfolios.[9]

[9]We exclude days with fewer than 10 firms in either portfolio leg.

We evaluate reduced SAE representations with k equal to 10, 50, 100, 500, 1,000, 3,000, and 5,000, as well as the full representation containing all 7,278 labeled features. For comparison, we extract standard final-layer embeddings from *Gemma-2-9B-IT* and use them as inputs to the same logistic regression and portfolio-construction procedure. This dense representation provides the benchmark against which we evaluate the predictive content of the sparse SAE features. Table IV reports the results.

Table IV. Return Predictions and Number of Features

This table reports the out-of-sample performance of return-prediction models based on dense final-layer embeddings and sparse SAE representations. The models are re-estimated annually using the preceding four years of data and predict the sign of the subsequent open-to-close return. Each day, firms are ranked by their predicted probabilities of a positive return. The long-short portfolios buy firms in the top 20% and short firms in the bottom 20%. *Benchmark* denotes the model using standard final-layer embeddings following Chen et al. (2022). The remaining columns report models using the top k economically labeled SAE features, with *All* denoting the full set of 7,278 features. *EW Sharpe* and *VW Sharpe* report the annualized Sharpe ratios of the equal- and value-weighted portfolios, respectively. *Accuracy* is the average daily percentage of firm-level predictions that match the sign of the realized return.

		Sparse features							
	Benchmark	10	50	100	500	1000	3000	5000	ALL
EW Sharpe	3.67	3.44	3.84	4.15	4.50	4.45	4.95	4.78	5.22
VW Sharpe	1.55	1.73	1.71	1.76	1.90	1.92	1.92	1.91	1.99
Accuracy	50.90%	50.92%	51.14%	51.33%	51.38%	51.33%	51.42%	51.38%	51.45%

The dense embedding benchmark produces annualized Sharpe ratios of 3.67 for the equal-weighted portfolio and 1.55 for the value-weighted portfolio, with an average daily accuracy of 50.90%. A sparse representation containing only 10 labeled features produces Sharpe ratios of 3.44 and 1.73, respectively, with similar predictive accuracy. With 50 labeled features, the SAE representation exceeds the dense benchmark under both weighting schemes, producing Sharpe ratios of 3.84 and 1.71. Performance is highest when all 7,278 labeled features are included. The full representation achieves Sharpe ratios of 5.22 and 1.99 and an average daily accuracy of 51.45%.

These results show that the SAE representation not only preserves the predictive content of dense embeddings but improves upon it. A representation containing only 50 economically labeled features already outperforms the dense benchmark under both weighting schemes. Performance generally rises as more labeled features are included and peaks with the full representation. Relative to the dense benchmark, the full set of 7,278 features raises the Sharpe ratio by approximately 42% for the equal-weighted portfolio and 28% for the value-weighted portfolio. Interpretability therefore comes with a substantial improvement, rather than a reduction, in predictive performance.

Decomposition of SAE Features We next use the SAE labels to identify the economic concepts represented among the features most relevant for predicting returns. For this descriptive analysis, we rank each feature by the average absolute magnitude of its fitted coefficient across the annual estimation windows. Figure 13 reports the composition of the top N features across the five economic themes, where N ranges from 10 to the full set of 7,278 labeled features. Appendix H provides a more granular decomposition into the 46 underlying topics.

The strongest individual predictors are concentrated in macroeconomic and market-wide concepts. Among the ten highest-ranked features, *Macro-Finance & Economics* and *Financial Markets & Instruments* each account for 40%, while *Corporate Finance* accounts for the remaining 20%. Neither *Investment & Portfolio Management* nor *Risk Management & Derivatives* appears among the top ten. As the representation expands, this concentration declines. The share of *Macro-Finance & Economics* falls from 40% to approximately 12%, and the share of *Financial Markets & Instruments* falls from 40% to approximately 19%. By contrast, *Corporate Finance* increases from 20% to nearly 30% and becomes the largest theme once N reaches approximately 1,000.

The result has a clear economic interpretation. A few macroeconomic and market-wide features apply broadly across firms and therefore dominate the top of the ranking. Corporate-finance information is more heterogeneous: individual concepts apply to narrower sets of firms, so its predictive content is distributed across a larger number of features. As the representation expands, its composition shifts from broad market signals toward a richer set of firm-specific corporate signals.

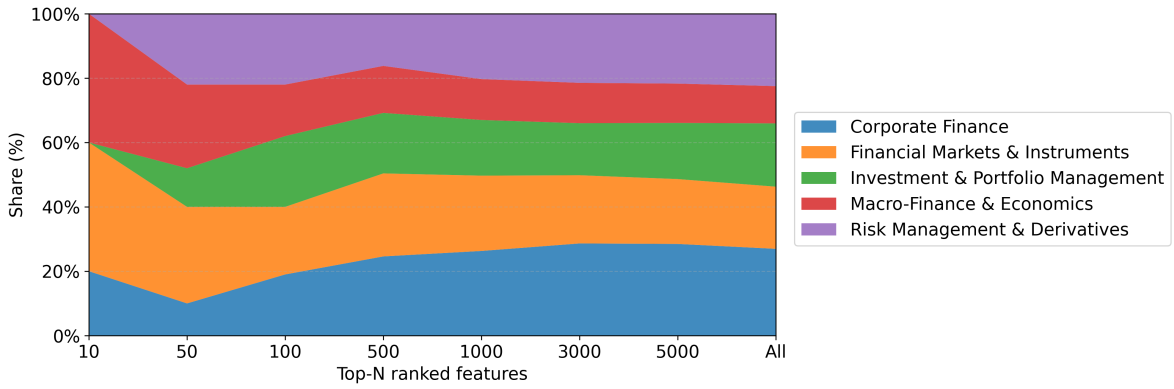


Figure 13. Economic Composition of the Most Predictive SAE Features
This figure decomposes the top N economically labeled SAE features across five economic themes. Features are ranked by the average absolute magnitude of their fitted coefficients across the annual estimation windows. The horizontal axis reports N, ranging from the 10 highest-ranked features to the full set of 7,278 labeled features. The vertical axis reports the percentage of the top N features assigned to each economic theme.

VII. Conclusion

This paper applies sparse autoencoders to make the internal representations of LLMs interpretable and controllable in economic applications. We develop a finance-specific labeling procedure that maps 7,278 SAE features to 46 economic topics organized into five broad themes. This representation reveals which economic concepts contribute to the model's outputs and allows researchers to intervene directly on selected concepts without retraining the underlying model. In our macroeconomic forecasting exercise, the concepts emphasized by the model vary systematically with the forecast target, even when the news context remains unchanged. The model's self-generated explanations, however, frequently diverge from these internal attributions. This result shows why interpreting an LLM through its stated reasoning alone can be misleading.

Our steering exercises demonstrate the practical importance of observing and controlling these internal representations. Prompt-based steering activates numerous collateral concepts and produces abrupt behavioral changes, making it difficult to attribute the resulting decisions to the intended treatment. SAE steering produces smoother changes in the targeted representation with substantially less collateral activation. In our applications, it preserves the model's selection of the higher-NPV project when risk aversion increases, corrects its flawed reasoning about Treasury bill rates during the financial crisis, and attenuates optimism bias in financial-news classification. The resulting news strategy outperforms the unsteered baseline, whereas prompt-based bias correction largely eliminates the model's predictive value. Finally, economically labeled SAE features match or exceed the performance

of dense final-layer embeddings in return prediction. Taken together, our results show that SAEs allow researchers to identify and steer the internal concepts that contribute to LLM outputs without sacrificing predictive performance.

REFERENCES

Aher, Gati, Rosa I. Arriaga, and Adam Tauman Kalai, 2023, Using large language models to simulate multiple humans and replicate human subject studies, in *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research* (PMLR).

Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate, 2023, Out of one, many: Using language models to simulate human samples, *Political Analysis* 31, 337–351.

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

Chen, Hui, Antoine Didisheim, and Luciano Somoza, 2024, Out of the (black) box: Ai as conditional probability, *Available at SSRN* .

Chen, Yifei, Bryan T. Kelly, and Dacheng Xiu, 2022, Expected returns and large language models, SSRN Working Paper No. 4543999.

Cheng, Qiang, Pengkai Lin, and Yue Zhao, 2024, Does generative ai facilitate investor trading? evidence from chatgpt outages, Available at SSRN 4872189.

Cook, Thomas R., Sophia Kazinnik, Zach Modig, and Nathan M. Palmer, 2026, What do LLMs want?, Finance and economics discussion series, Board of Governors of the Federal Reserve System.

Cunningham, Hoagy, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey, 2023, Sparse autoencoders find highly interpretable features in language models, *arXiv preprint arXiv:2309.08600* .

Didisheim, Antoine, Martina Fraschini, and Luciano Somoza, 2025a, AI's predictable memory and its implications for finance, Available at SSRN 5331177.

Didisheim, Antoine, Martina Fraschini, and Luciano Somoza, 2025b, Ai's predictable memory and its implications for finance, SSRN Working Paper No. 5331177.

Didisheim, Antoine, Bryan T. Kelly, Mohammad Pourmohammadi, and Hanqing Tian, 2026, The inefficient pricing of news, NBER Working Paper 35093, National Bureau of Economic Research.

Engelberg, Joseph, Asaf Manela, William Mullins, and Luka Vulicevic, 2025, Entity neutering, Available at SSRN 5182756.

Engelberg, Joseph E, and Christopher A Parsons, 2011, The causal impact of media in financial markets, *the Journal of Finance* 66, 67–97.

Fedyk, Anastassia, Ali Kakhbod, Peiyao Li, and Ulrike Malmendier, 2024, Chatgpt and perception biases in investments: An experimental study, Available at SSRN 4787249.

Fieberg, Christian, Lars Hornuf, Maximilian Meiler, and David Streich, 2025, Using large language models for financial advice, CESifo Working Paper 11666, CESifo.

Gallegos, Isabel O., Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed, 2024, Bias and fairness in large language models: A survey, *Computational Linguistics* 50, 1097–1134.

Gao, Leo et al., 2024, Scaling and evaluating sparse autoencoders, *arXiv preprint arXiv:2406.04093* .

Gemini, 2025, Gemini: A family of highly capable multimodal models.

Giglio, Stefano, Bryan Kelly, and Dacheng Xiu, 2022, Factor models, machine learning, and asset pricing, *Annual Review of Financial Economics* 14, 337–368.

Goodfellow, Ian, Yoshua Bengio, Aaron Courville, and Yoshua Bengio, 2016, *Deep learning*, volume 1 (MIT press Cambridge).

Gu, Shihao, Bryan Kelly, and Dacheng Xiu, 2020, Empirical asset pricing via machine learning, *The Review of Financial Studies* 33, 2223–2273.

Gu, Shihao, Bryan T. Kelly, and Dacheng Xiu, 2021, Autoencoder asset pricing models, *Journal of Econometrics* 222, 429–450.

Horton, John J., 2023, Large language models as simulated economic agents: What can we learn from homo silicus?, NBER Working Paper 31122, National Bureau of Economic Research.

Kelly, Bryan, Semyon Malamud, and Kangying Zhou, 2024, The virtue of complexity in return prediction, *The Journal of Finance* 79, 459–503.

Kim, Alex, Maximilian Muhn, and Valeri Nikolaev, 2024a, Financial statement analysis with large language models, *arXiv preprint arXiv:2407.17866* .

Kim, Jeongbin, Matthew Kovach, Kyu-Min Lee, Euncheol Shin, and Hector Tzavellas, 2024b, Can an LLM learn preferences from choice data?, *arXiv preprint arXiv:2401.07345* .

Korinek, Anton, 2023, Generative ai for economic research: Use cases and implications for economists, *Journal of Economic Literature* 61, 1281–1317.

Lieberum, Tom, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda, 2024, Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2, *arXiv preprint arXiv:2408.05147* .

Liu, Jiaxin, Yixuan Tang, Yi Yang, and Kar Yan Tam, 2025, Evaluating and aligning human economic risk preferences in LLMs, *arXiv preprint arXiv:2503.06646* .

Liu, Nicole, and Sam Kirshner, 2024, The futures too bright: ChatGPT's optimism forecasting bias, in *Proceedings of the 45th International Conference on Information Systems (ICIS 2024)*, Bangkok, Thailand.

Llama, 2024, The llama 3 herd of models.

Lopez-Lira, Alejandro, and Yuehua Tang, 2023, Can chatgpt forecast stock price movements? return predictability and large language models, arXiv preprint arXiv:2304.07619.

Lopez-Lira, Alejandro, Yuehua Tang, and Mingyin Zhu, 2025, The Memorization Problem: Can We Trust LLMs' Economic Forecasts?, arXiv preprint arXiv:2504.14765.

Loughran, Tim, and Bill McDonald, 2011, When is a liability not a liability? textual analysis, dictionaries, and 10-ks, *The Journal of finance* 66, 35–65.

Ludwig, Jens, Sendhil Mullainathan, and Ashesh Rambachan, 2025, Large language models: An applied econometric framework, Technical report, National Bureau of Economic Research.

OpenAI, 2024, Gpt-4 technical report.

Park, Joon Sung, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein, 2023, Generative agents: Interactive simulacra of human behavior, in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)* (ACM).

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever, 2019, Language models are unsupervised multitask learners .

Sarkar, Suproteem K., and Keyon Vafa, 2024, Lookahead bias in pretrained language models, SSRN Working Paper No. 4754678.

Shi, Wei et al., 2025, Route sparse autoencoder to interpret large language models, *arXiv preprint arXiv:2503.08200* .

Tetlock, Paul C, 2007, Giving content to investor sentiment: The role of media in the stock market, *The Journal of finance* 62, 1139–1168.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin, 2017, Attention is all you need.

Zhang, Yizhong, and Guofu Zhou, 2026, Large language models for asset pricing: Learning from earnings calls, *Available at SSRN 6712298* .

Appendix A. Technical Details of Sparse Autoencoders

This appendix provides the formal counterpart to the methodology summarized in Section II, covering the language-model and sparse-autoencoder setup, the training objective, the feature-labeling and selection procedure, and the mechanics of concept steering.

The technology underlying generative models like ChatGPT and similar LLMs, is based on *generative architectures* that model the probability distribution of text sequences. At their core, these models estimate the conditional probability of the next token, typically a word or subword, given the preceding sequence of text.

These generative models rely on stacked *transformer blocks*. Each block applies a complex transformation to the output of the previous block (the model's internal state) and passes the updated state to the next layer. This layered design lets the model gradually learn more complex language patterns as information moves up through the layers. (see, e.g Vaswani et al., 2017; Radford et al., 2019).

Formally,[10] let us denote the internal state of the model at layer l as $\mathbf{r}^{(l)}$, a vector in $\mathbb{R}^d$ for some d. This vector, often called the *residual stream*, serves as a running summary of the information accumulated up to that point in the sequence. Each transformer block, denoted by $\text{Block}^{(l)}$, receives the current residual stream, applies a learned transformation, and provides a modification to the stream. The architecture is structured such that this modification is added back to the original input:

$$\mathbf{r}^{(l+1)} = \mathbf{r}^{(l)} + \text{Block}^{(l)}(\mathbf{r}^{(l)}). \tag{A1}$$

This recursive formulation captures the essential mechanism, where each layer refines the current representation by processing and reintegrating new information. With m total blocks, the final residual stream $\mathbf{r}^{(m)}$ is passed to an output module that maps it to a probability distribution over the vocabulary. This distribution represents the model's prediction for the next word:

$$P(\text{Next Word} \mid \text{Input Text}) = \text{Block}^{(\text{Output})}(\mathbf{r}^{(m)}). \tag{A2}$$

This architecture, shared by many state-of-the-art models, has demonstrated remarkable performance across a wide range of NLP benchmarks (e.g., OpenAI, 2024; Llama, 2024;

[10] For simplicity, we omit the input length dimension and present the model as if processing a single token at a time. This simplification does not affect the core intuition and is used purely for notational convenience.

Gemini, 2025), yet it remains notoriously difficult to interpret. The source of this opacity is twofold. First, the individual transformation blocks are large, nonlinear, and densely parameterized, which makes them uninterpretable to a human. Second, the residual streams are dense. Individual concepts like "optimism" or "pessimism" are unlikely to be captured by a single point on the stream. Instead, individual concepts are captured by a complex combination of many individual points. In the human brain, different regions often specialize in distinct tasks such as language, vision, or motor control, making it easier to associate structure with function. Large language models, by contrast, have no such architectural or evolutionary incentive to develop specialized pathways. Their sole training objective is to predict the next token, which can be met using highly distributed representations with no clear division of labor across dimensions.

This lack of interpretability has been a persistent concern, especially in high-stakes or domain-sensitive applications (Ludwig et al., 2025). Without insight into what information is stored in $\mathbf{r}^{(l)}$ or how it evolves across layers, understanding how LLMs process financial information or tweaking this process is virtually impossible.

To address this, we seek to discover a sparse and semantically meaningful representation of the residual stream. Specifically, we aim to identify a transformation:

$$\mathbf{r}^{(l)} \mapsto \mathbf{z}^{(l)} \in \mathbb{R}^k$$

where each coordinate of $\mathbf{z}^{(l)}$ corresponds to an *interpretable* feature or concept. Intuitively, this would enable us to "open the black box" by exposing a layer-wise semantic decomposition of the model's internal reasoning.

A. Sparse Autoencoders (SAEs)

Autoencoders are a class of neural network models designed for unsupervised dimensionality reduction. Like Principal Component Analysis (PCA), they aim to compress input data into a lower-dimensional latent representation that minimizes the reconstruction error. However, whereas PCA minimizes reconstruction error using linear projections onto orthogonal principal components, autoencoders directly minimize the reconstruction error without imposing linearity or orthogonality constraints, allowing them to learn more flexible and potentially more powerful nonlinear mappings. Their ability to learn nonlinear embeddings has made them a powerful tool for data compression, feature extraction, and anomaly detection. In finance, autoencoders have been applied to problems such as factor discovery and risk modeling (see, e.g., Gu et al., 2021).

Formally, an autoencoder consists of two components: an encoder and a decoder. Given an input vector $\mathbf{x} \in \mathbb{R}^d$, the encoder maps it to a latent representation $\mathbf{z} \in \mathbb{R}^k$ via a learned transformation:

$$\mathbf{z} = f(\mathbf{W}_e \mathbf{x} + \mathbf{b}_e), \tag{A3}$$

where $\mathbf{W}_e \in \mathbb{R}^{k \times d}$ is the encoder weight matrix, $\mathbf{b}_e \in \mathbb{R}^k$ is a bias vector, and $f(\cdot)$ is a nonlinear activation function.[11] For simplicity, we present here the case of a single-layer encoding and decoding model. However, as with any neural network, these layers can be stacked on top of one another.

The decoder then attempts to reconstruct the original input from $\mathbf{z}$:

$$\hat{\mathbf{x}} = g(\mathbf{W}_d \mathbf{z} + \mathbf{b}_d), \tag{A4}$$

where $\mathbf{W}_d \in \mathbb{R}^{d \times k}$ and $\mathbf{b}_d \in \mathbb{R}^d$ are decoder parameters, and $g(\cdot)$ is usually chosen to be the identity function or a sigmoid, depending on the domain of the input data (Goodfellow et al., 2016).

The encoder and decoder are trained jointly to minimize the reconstruction error between the input $\mathbf{x}$ and its reconstruction $\hat{\mathbf{x}}$:

$$\mathcal{L}_{\text{AE}}(\mathbf{x}) = \underbrace{\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2}_{\text{Reconstruction loss}} . \tag{A5}$$

Autoencoders are powerful because they can compress data into a lower-dimensional representation, addressing one of the two sources of opacity we discussed earlier: the sheer size and complexity of the transformation blocks. However, they do not address the other major challenge: in order to be interpretable, this new, smaller representation has to be sparse. This is where *Sparse Autoencoders (SAEs)* can play a role. A SAE is a variant of the standard autoencoder that adds a sparsity-inducing penalty, typically an ℓ_1 or ℓ_0 norm on the latent representation $\mathbf{z}$. The aim is to encourage most dimensions of $\mathbf{z}$ to be inactive (conceptually zero for any given input), thereby promoting interpretability and disentanglement in the learned features:

$$\mathcal{L}_{\text{SAE}}(\mathbf{x}) = \underbrace{\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2}_{\text{Reconstruction}} + \underbrace{\lambda \|\mathbf{z}\|_1}_{\text{Sparsity penalty}} , \tag{A6}$$

[11] While k is typically smaller than d in standard autoencoders, in the context of LLM interpretability, it is common to set $k > d$ to allow for overcomplete, disentangled representations and lower reconstruction error.

where $\lambda > 0$ controls the strength of the sparsity constraint.

This sparse structure is particularly attractive for interpretability: if only a small number of latent features are active for a given residual stream $\mathbf{x}$, then those features can be more easily inspected and understood in isolation.

B. Training Sparse Autoencoders within a Large Language Model

The concept of training Sparse Autoencoders (SAEs) within a Large Language Model (LLM) to enhance interpretability was introduced by Cunningham et al. (2023).

In this and the following section, we describe how SAEs can be used to uncover semantically meaningful structures in the internal representations of LLMs. As illustrated in Figure A1, the objective is to construct an augmented model pipeline in which, at each layer l, the residual stream $\mathbf{r}^{(l)}$ is mapped to a sparse code $\mathbf{z}^{(l)}$. The individual dimensions of $\mathbf{z}^{(l)}$ are designed to correspond to interpretable concepts or functions.

To achieve this, SAEs are trained on the residual streams produced at a fixed layer l in the language model. More precisely, an input text is run through the LLM up to layer l. The corresponding residual stream vector (i.e., the current internal representation of the input at a given layer in the model's processing) $\mathbf{r}^{(l)} \in \mathbb{R}^d$ is then recorded. Repeating this process across a large corpus gives a dataset of residual streams:

$$\mathcal{D} = \left\{ \mathbf{r}_i^{(l)} \right\}_{i=1}^{N},$$

where each $\mathbf{r}_i^{(l)}$ is the residual stream at layer l for the i-th text input.

A sparse autoencoder is then trained to compress and reconstruct each $\mathbf{r}_i^{(l)}$ through a sparse latent representation:

$$\mathbf{z}_i^{(l)} = f(\mathbf{W}_e \mathbf{r}_i^{(l)} + \mathbf{b}_e), \tag{A7}$$

$$\hat{\mathbf{r}}_i^{(l)} = g(\mathbf{W}_d \mathbf{z}_i^{(l)} + \mathbf{b}_d), \tag{A8}$$

where $f(\cdot)$ is a nonlinear activation function, and $g(\cdot)$ is usually the identity.

The loss function encourages both accurate reconstruction and sparsity in the hidden representation:

$$\mathcal{L}(\mathbf{r}_i^{(l)}) = \underbrace{\|\mathbf{r}_i^{(l)} - \hat{\mathbf{r}}_i^{(l)}\|_2^2}_{\text{Reconstruction loss}} + \lambda \underbrace{\|\mathbf{z}_i^{(l)}\|_1}_{\text{Sparsity penalty}}, \tag{A9}$$

with $\lambda > 0$ controlling the trade-off between fidelity and interpretability.

Since the autoencoder is a neural network without a closed-form solution, it is trained by minimizing the expected loss across the generated dataset of residual streams.[12]

$$\min_{\Theta} \mathbb{E}_{\mathbf{r}\sim\mathcal{D}}[\mathcal{L}(\mathbf{r})], \tag{A10}$$

One core advantage of the SAEs is that it does not require retraining the LLM itself. Instead, the SAE is trained independently on the residual stream of a pre-trained model. This makes the method computationally efficient and accessible, requiring only a fraction of the data and computational resources typically needed to train or fine-tune a full-scale LLM.

In practice, the SAE can be viewed as an interpretable "plug-in" that augments the internal representations of a transformer model without altering its original parameters. This separation is particularly advantageous in settings where full retraining is impractical or undesirable, such as applied research in finance. Moreover, the increasing availability of open-source models and pretrained SAEs makes this approach highly suitable for academic and applied analysis.

Throughout this paper, we utilize the open-source Sparse Autoencoders released by Google DeepMind (Lieberum et al., 2024), trained on intermediate representations from the instruction-tuned language model Gemma-2-9B-IT. These models provide an ideal testbed for our study, combining open accessibility with strong performance on language understanding tasks.

C. Feature Labeling and Selection

While SAEs provide a sparse and interpretable representation of the model's internal states, the features themselves do not come with human-interpretable meanings. The most common labeling approach is to collect the tokens or text passages that most strongly activate each feature and ask a language model to summarize the shared theme (e.g., Cunningham et al., 2023). Publicly available resources such as Neuronpedia provide large collections of such feature annotations derived from generic corpora.[13] However, because these labels are inferred from general-purpose text, many features correspond to concepts unrelated to finance or economics, ranging from cooking recipes to sports events, making it difficult to directly connect the feature space to economic meanings.

We instead label features from the other direction: we define the concepts we care about

[12]These residual streams are obtained from forward passes of the pretrained language model over text corpora, such as books, news articles, and web pages, similar to those used in LLM training but typically drawn from a smaller, curated subset.

[13]See https://www.neuronpedia.org/

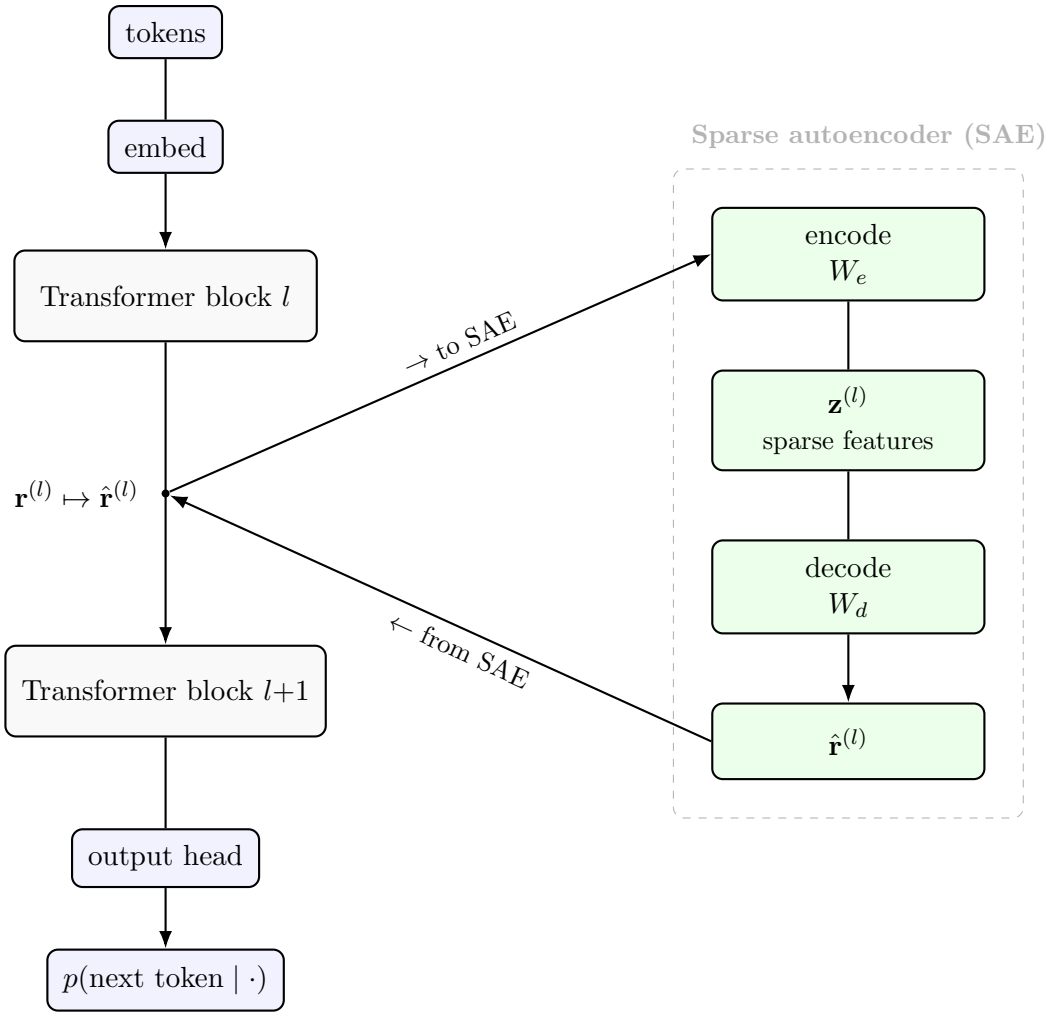


Figure A1. Illustration of SAE's integration with a Transformer model.
The residual stream $\mathbf{r}^{(l)}$ is extracted from an intermediate layer of the language model and projected onto a sparse, interpretable representation $\mathbf{z}^{(l)}$. This sparse code is then decoded to reconstruct $\hat{\mathbf{r}}^{(l)}$, which is fed back into the model for subsequent processing.

upfront, and find which features respond to them. Concretely, we specify $S = 46$ economic topics (e.g., *Leverage*, *Monetary Policy*, *Risk Aversion*), grouped into five main themes: *Macro-Finance & Economics*, *Financial Markets & Instruments*, *Corporate Finance*, *Investment & Portfolio Management*, and *Risk Management & Derivatives*. For each topic i, we use GPT to generate a set of representative example texts and pass them through the SAE-augmented LLM, recording the mean activation of each feature. Prompt 9 reports the exact instruction used to generate them.

This produces an activation matrix $\bar{z} \in \mathbb{R}^{S \times k}$, where $\bar{z}_{i,j}$ is the average SAE activation of feature j across all example texts for topic i.

To measure how *exclusively* feature j responds to topic i, as opposed to firing uniformly

```
Act as a financial journalist.

Generate exactly {phrases_per_run} short examples that a financial
journalist might use when writing about "{small_topic}" under the broader
 topic of "{big_category}".

Constraints:
- Each example must be a short phrase, clause, or sentence fragment.
```

Prompt 9. Financial Journalism Phrase Generation Prompt
This prompt asks the model to act as a financial journalist and generate a fixed number of short examples related to a specified small topic within a broader financial category. The examples are constrained to be brief phrases, clauses, or sentence fragments.

across all topics, we center each feature's activation profile across topics and clip negatives:

$$\tilde{z}_{i,j} \;=\; \max\!\left(0,\; \bar{z}_{i,j} - \tfrac{1}{S}\sum_{i'} \bar{z}_{i',j}\right). \tag{A11}$$

Normalizing column-wise across topics gives a *probability map*:

$$\pi_{i,j} \;=\; \frac{\tilde{z}_{i,j}}{\sum_{i'} \tilde{z}_{i',j}}, \qquad \sum_{i} \pi_{i,j} = 1, \tag{A12}$$

so $\pi_{i,j}$ is the probability that feature j "belongs to" topic i: a feature receives positive mass only for topics where it activates above its mean level across all topics.

We then assign each feature to the topic for which it is both strongly active and highly specific, via the combined score

$$s_{i,j} \;=\; \bar{z}_{i,j} \cdot \pi_{i,j}, \tag{A13}$$

which rewards features that are both strongly active and concentrated on topic i. Each feature is hard-assigned to its highest-scoring topic:

$$\hat{i}(j) \;=\; \arg\max_{i}\; s_{i,j}. \tag{A14}$$

To quantify the overall concentration of a feature across topics, we compute the normalized entropy of $\pi_{\cdot,j}$:

$$H_j \;=\; -\frac{1}{\log S}\sum_{i=1}^{S} \pi_{i,j} \log \pi_{i,j} \;\in\; [0,1], \tag{A15}$$

with $H_j = 0$ for a feature that activates exclusively for one topic and $H_j = 1$ for uniform activation.

Feature selection. Not all features assigned to a topic are equally informative. For each topic i, we form a candidate pool of features that are exclusively assigned to it and have sufficiently low entropy ($H_j < \tau_H = 0.6$):[14]

$$\mathcal{C}_i \;=\; \big\{ j : \hat{i}(j) = i \;\text{ and }\; H_j < \tau_H \big\}. \tag{A16}$$

Within $\mathcal{C}_i$, features are ranked by score $s_{i,j}$ in descending order, and we greedily include features until their cumulative score mass reaches a fraction $\tau_c = 0.8$ of the total scores, and $\mathcal{F}_i$ will be the final set of features we assign to topic i:

$$\mathcal{F}_i \;=\; \text{top-}n_i^* \text{ features in } \mathcal{C}_i, \quad n_i^* = \min\left\{ n : \frac{\sum_{r=1}^{n} s_{i,(r)}}{\sum_{j \in \mathcal{C}_i} s_{i,j}} \geq \tau_c \right\}, \tag{A17}$$

where $s_{i,(r)}$ is the r-th largest score in $\mathcal{F}_i$. This selection criterion is self-calibrating: topics dominated by a single feature select a small set, while diffuse topics select more, always capturing a consistent fraction of the topic's total signal. At the same time, by truncating the tail of low-scoring features, it excludes those with negligible contribution, effectively filtering out weak, noisy activations that would otherwise dilute the semantic coherence of the topic.

D. Concept Steering

The labeling approach above assigns each feature to an economic topic based on which example texts activate it. A natural question is whether these assignments are meaningful: does steering the model with a feature labeled *risk-aversion* actually cause it to be more risk-averse? Concept steering provides a direct answer.

At a high level, steering can be understood as follows:

1. Identify a relevant feature in the sparse representation to steer the LLM toward. For instance, selecting one associated with risk to promote risk-averse behavior.
2. Use the SAE's decoder to identify the perturbation induced by activating this feature; specifically, what would have been *added* to the residual stream if the feature were more strongly activated.

[14] This corresponds to a feature distributing its mass over roughly $\exp(H_j \log S)$ topics; with $S = 46$ and $H_j = 0.6$, this is on the order of $\sim$10 effective topics.

3. Add this perturbation to the residual stream to obtain a steered probability distribution over the next tokens.
4. Iteratively sample tokens from this steered distribution to generate text aligned with the intended concept (e.g., more risk-averse responses).

Formally, recall from Equation (A2) that the probability of the next token in an LLM is obtained by applying the final block to the last residual stream:

$$P(\text{Next Word} \mid \text{Input Text}) = \text{Block}^{(\text{Output})}(\mathbf{r}^{(m)}).$$

Assume an SAE has been trained on the m^{th} residual stream, with encoder[15]

$$\mathbf{z}^{(m)} = f(\mathbf{W}_e \mathbf{r}^{(m)} + \mathbf{b}_e),$$

and decoder

$$\hat{\mathbf{r}}^{(m)} = g(\mathbf{W}_d \mathbf{z}^{(m)} + \mathbf{b}_d).$$

Suppose a relevant feature $z_j^{(m)}$ of the sparse representation has been identified, for example, the topic feature assigned under risk-aversion. Steering then consists of generating a perturbation vector

$$\mathbf{e}_j = g(\mathbf{W}_d \mathbf{S}_j^{(m)} + \mathbf{b}_d),$$

where $\mathbf{S}_j^{(m)}$ is a steering vector that is zero everywhere except at the j^{th} coordinate, which is assigned a nonzero steering value $s \neq 0$:

$$\mathbf{S}_j^{(m)} = [\,0, \ldots, 0, \underbrace{s}_{j^{th} \text{ position}}, 0, \ldots, 0\,]. \tag{A18}$$

The probability of the next token under steering is then

$$P^{(\text{Steered})}(\text{Next Word} \mid \text{Input Text}) = \text{Block}^{(\text{Output})}(\mathbf{r}^{(m)} + \mathbf{e}_j).$$

We show an example of steering to validate our labeling approach in Section II.B.

This approach provides a fine-grained, concept-level control mechanism grounded in interpretable internal features. Unlike prompt engineering, which relies on heuristic text manipulations and often lacks transparency, steering via sparse features operates directly on the model's latent representations. This control enables systematic and reproducible inter-

[15] In practice, SAE can be trained on any intermediary layer, but for the sake of clarity we focus on this simplified example. The same logic applies if the perturbation occurs earlier in the LLMs.

ventions without altering the model's parameters or retraining.

Appendix B. Topic Word Clouds

This appendix reports the word clouds for all 46 topics, extending the three examples shown in Figure 3. We steer the features assigned to each topic to their maximum activation and ask the model to complete Prompt 1, holding the prompt fixed across topics. Figure A2 collects the four panels. The generated vocabulary matches the assigned label in nearly every topic, which confirms that the labeling pipeline recovers the intended economic concepts.

Appendix C. Topic Decomposition of Macroeconomic Expectations

This appendix reports the mechanistic topic decomposition for all fourteen macroeconomic variables summarized in Section III.D. For each variable, we average the monthly feature contribution over the sample and aggregate it to the five main themes and their constituent topics. Figure A3 collects the four panels. The dominant theme and its leading topics vary systematically across forecast targets, even though the news context is identical across variables.

Appendix D. Time Series of Feature Contributions

This appendix reports, for each of the fourteen macroeconomic variables, the monthly contribution of every main theme to the LLM's forecast over the 2000–2021 sample. Each contribution is the monthly measure $\bar{c}^{v}_{j,t}$ aggregated to the five main themes, plotted in absolute value so that the height of each band reflects the magnitude with which the theme drives the forecast in a given month. These series underpin the dynamic validation discussed in Section III.E. Figures A4 through A17 report one variable each. For most variables a single theme dominates the forecast, and its contribution peaks around the episodes that made that theme salient.

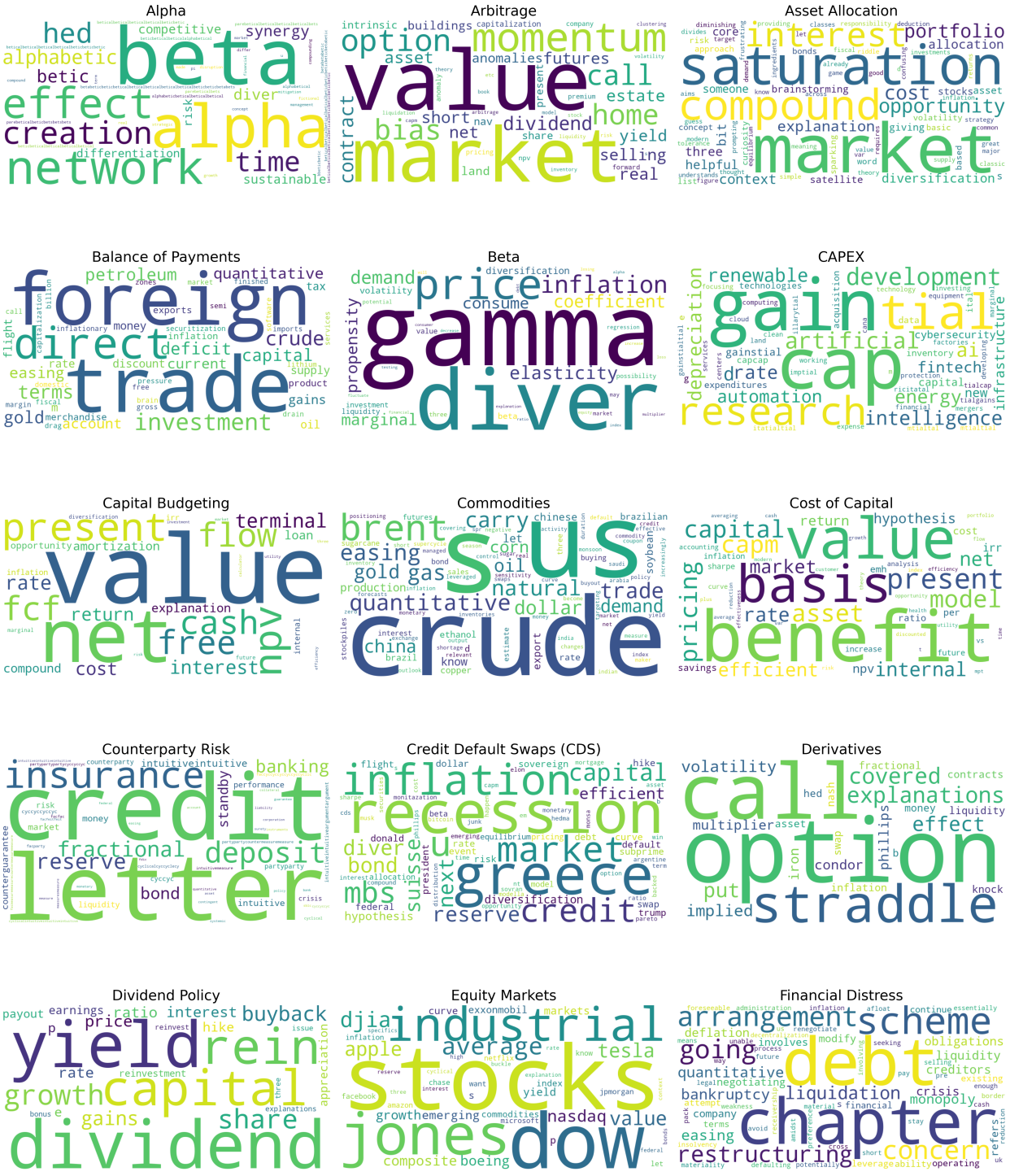


Figure A2. Word clouds for steered SAE topics.
Each block within the panel shows the most frequent words generated by the LLM when its residual stream is steered along the top features assigned to a given topic, in response to Prompt 1. Font size is proportional to frequency.

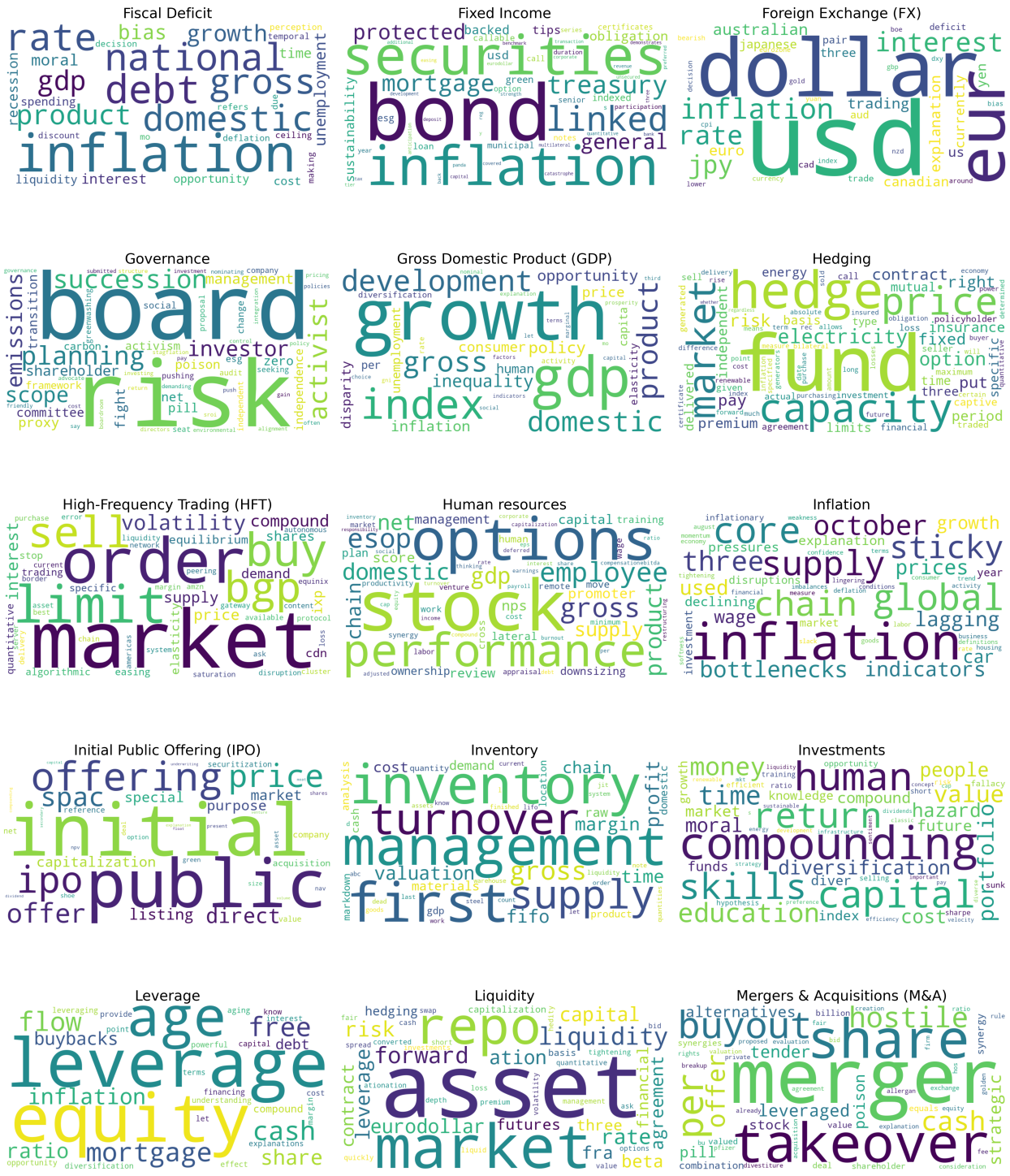


Figure A2. Word clouds for steered SAE topics, continued.



Figure A2. Word clouds for steered SAE topics, continued.

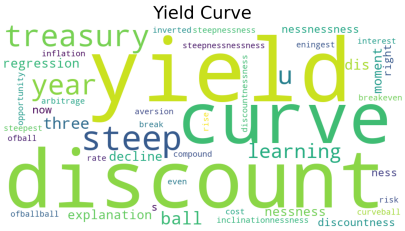


Figure A2. Word clouds for steered SAE topics, continued.

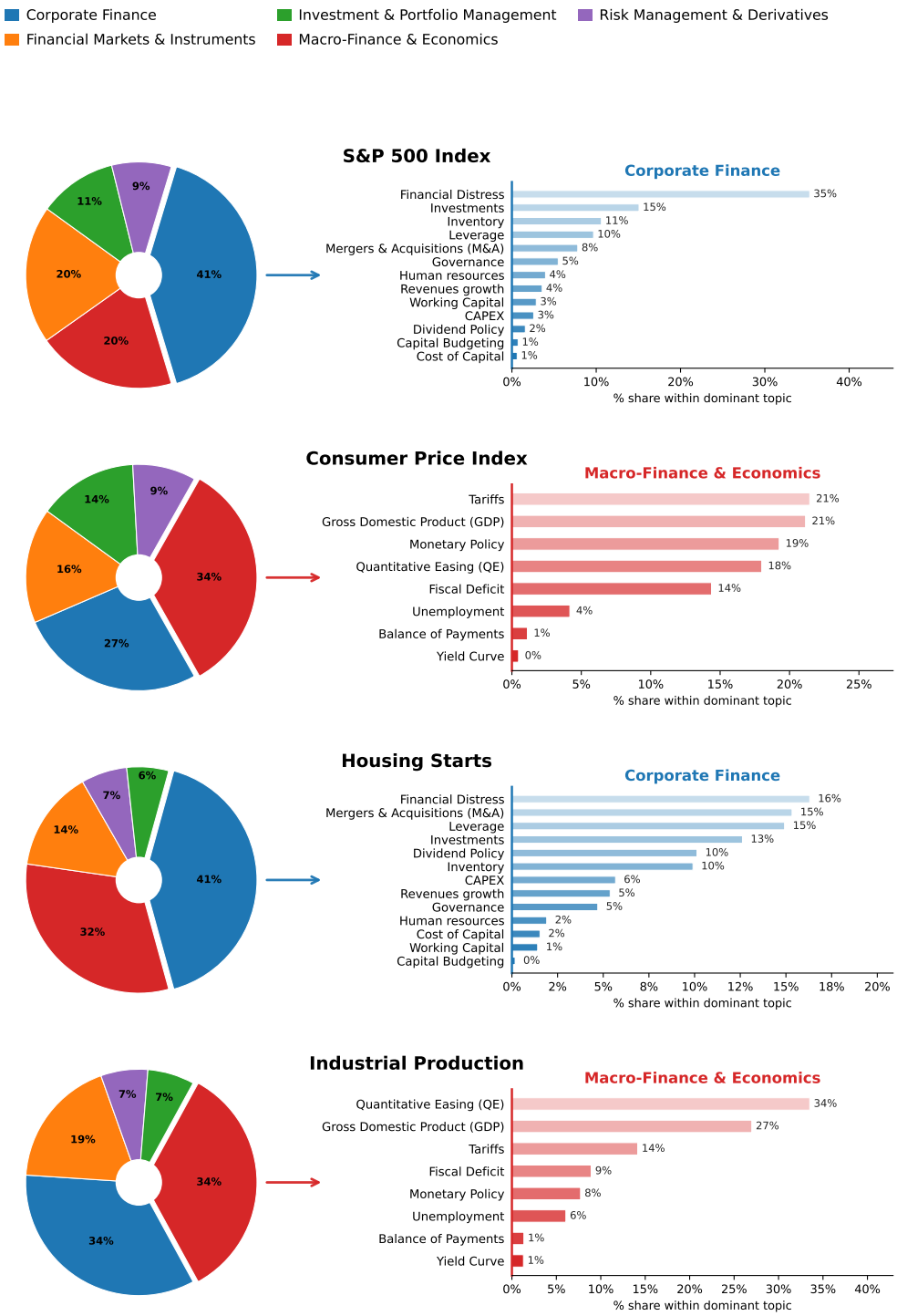


Figure A3. Decomposition of topic contributions

For each variable in the panel, the pie chart on the left reports the share of total feature-contribution mass accounted for by each of the five main themes (*Macro-Finance & Economics*, *Financial Markets & Instruments*, *Corporate Finance*, *Investment & Portfolio Management*, *Risk Management & Derivatives*), and the bar chart on the right decomposes the dominant theme into its constituent topics.

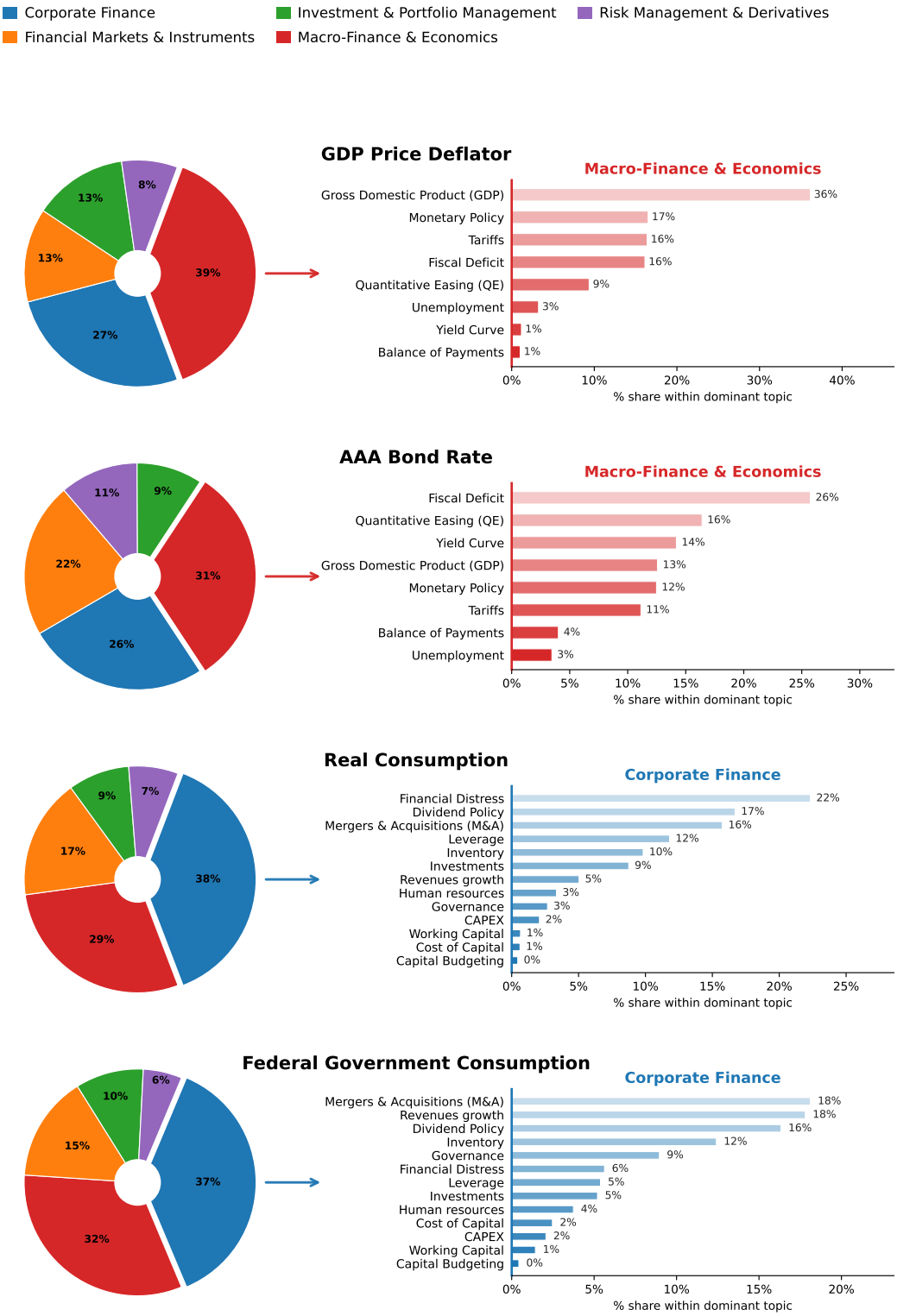


Figure A3. Decomposition of topic contributions, continued

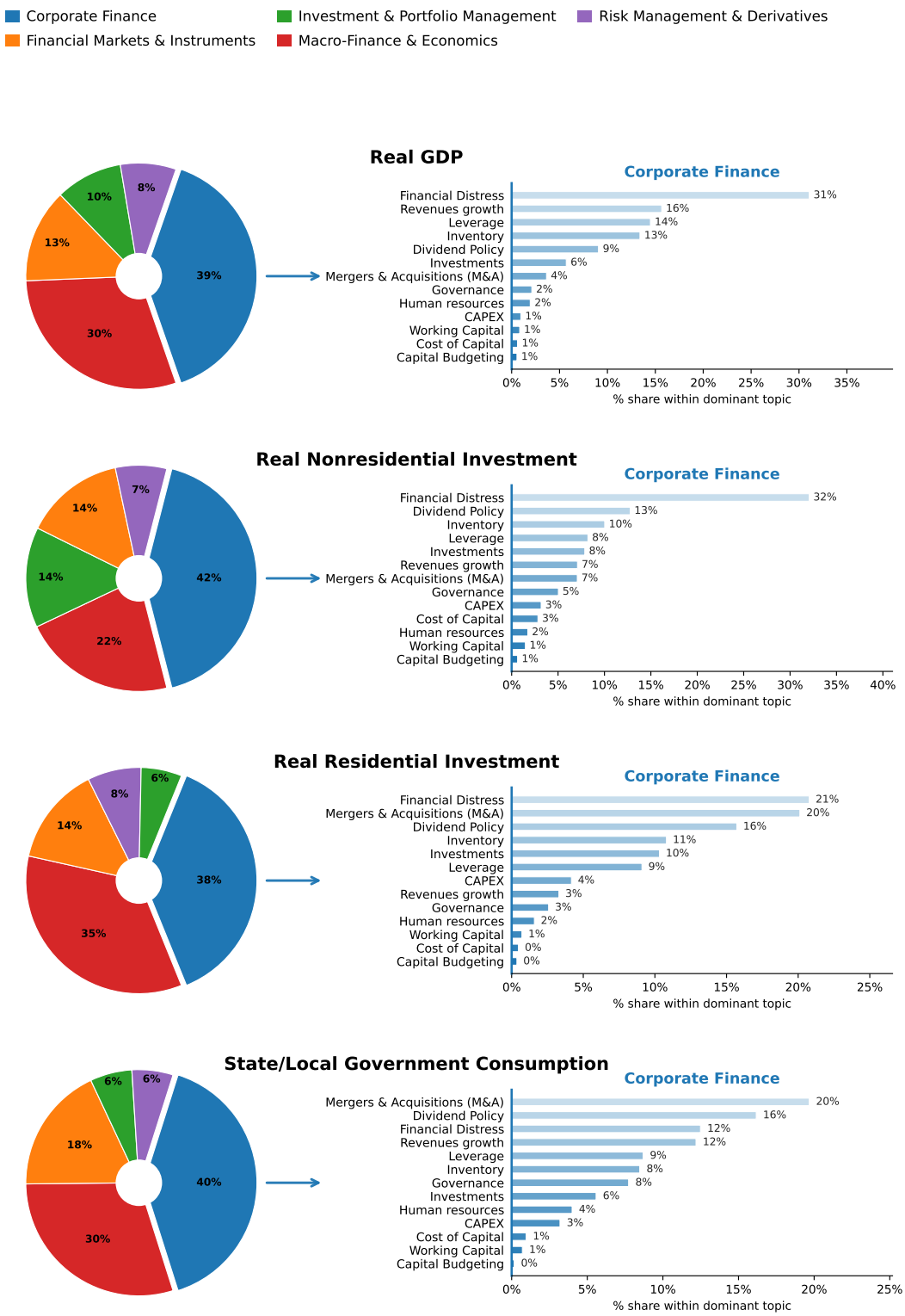


Figure A3. Decomposition of topic contributions, continued

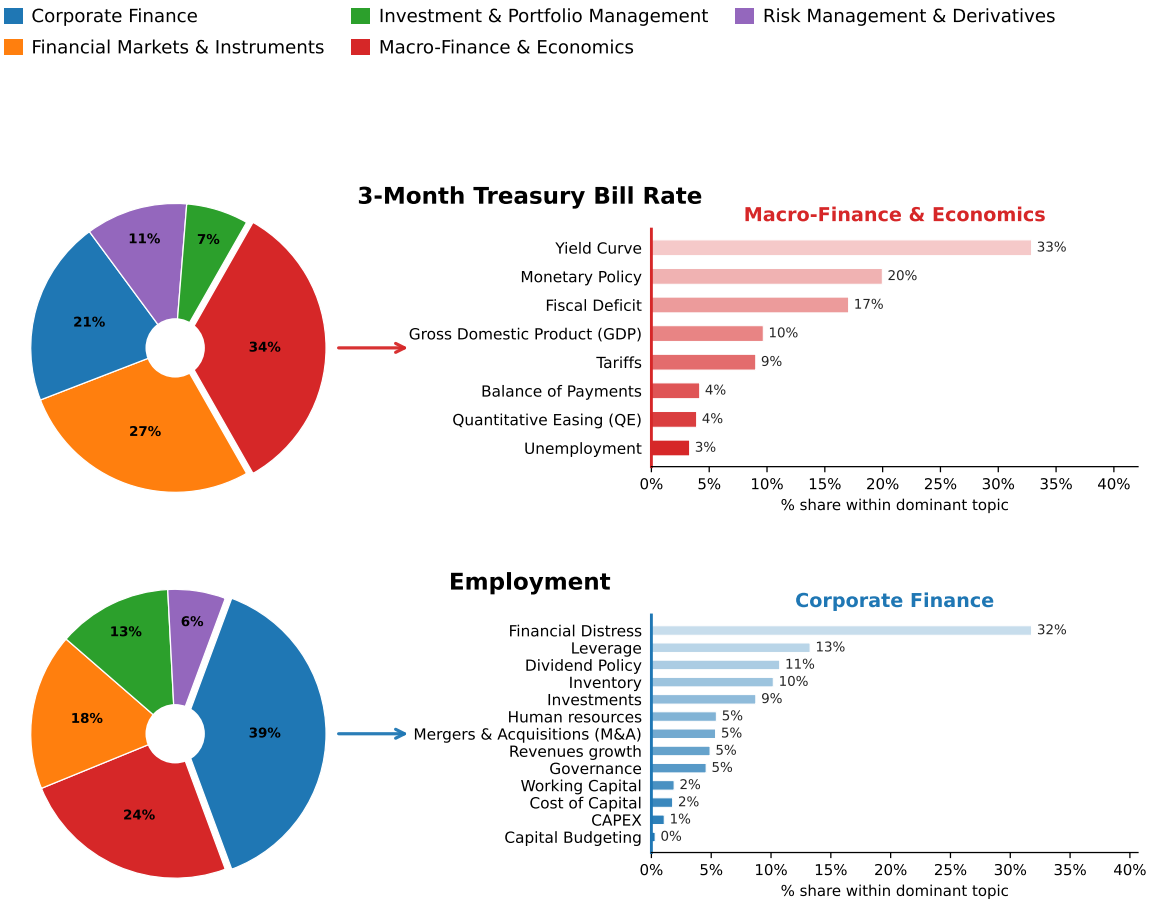


Figure A3. Decomposition of topic contributions, continued

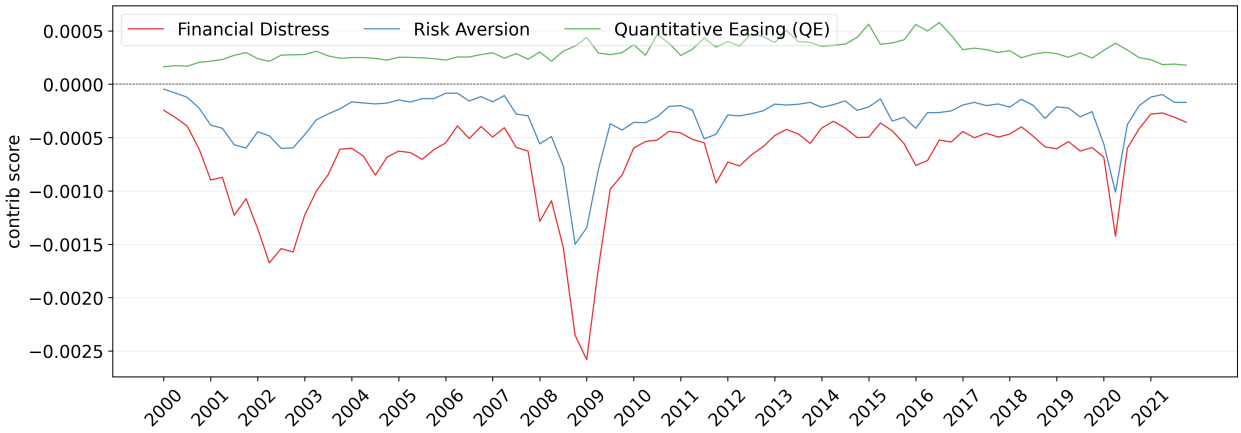


Figure A4. Time series of theme contributions: S&P 500 Index.

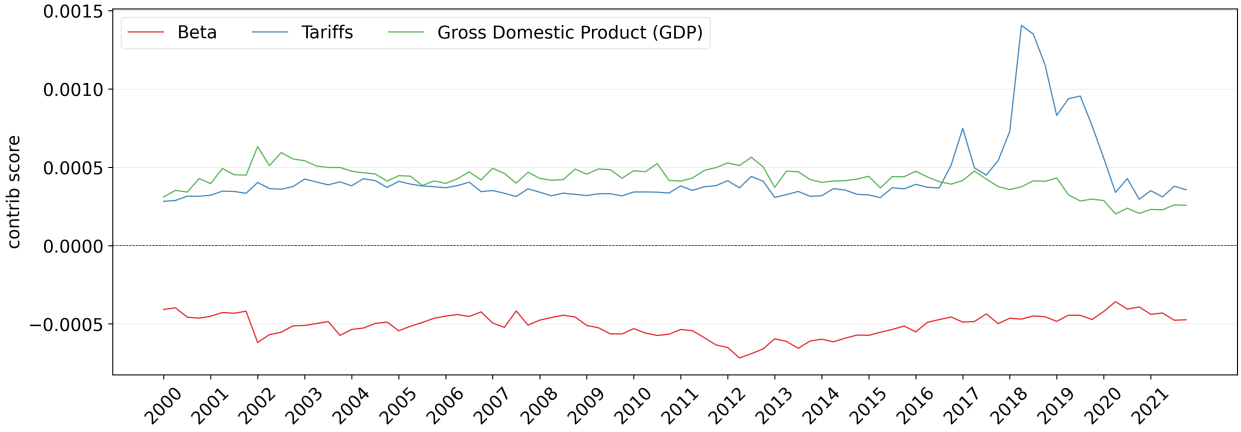


Figure A5. Time series of theme contributions: Consumer Price Index.

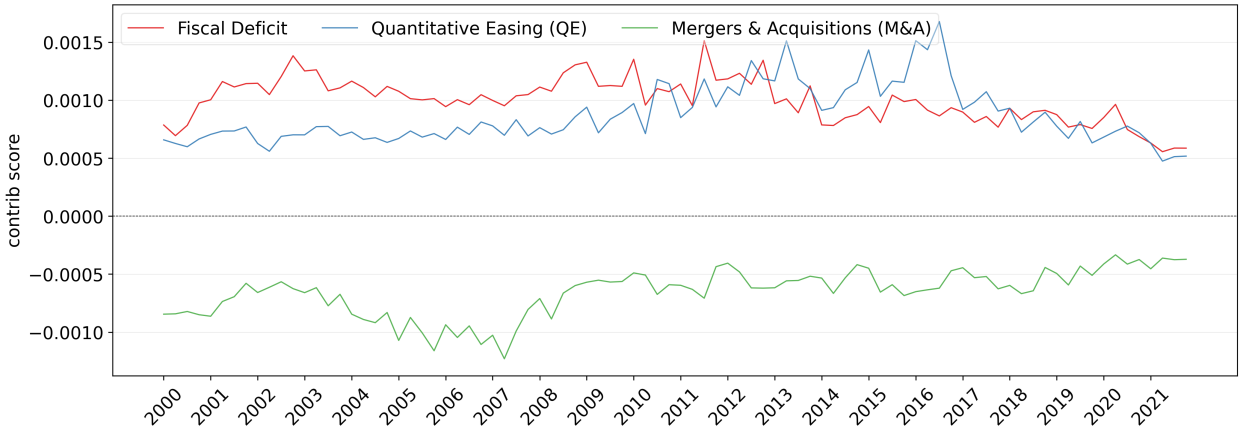


Figure A6. Time series of theme contributions: Federal Government Consumption.

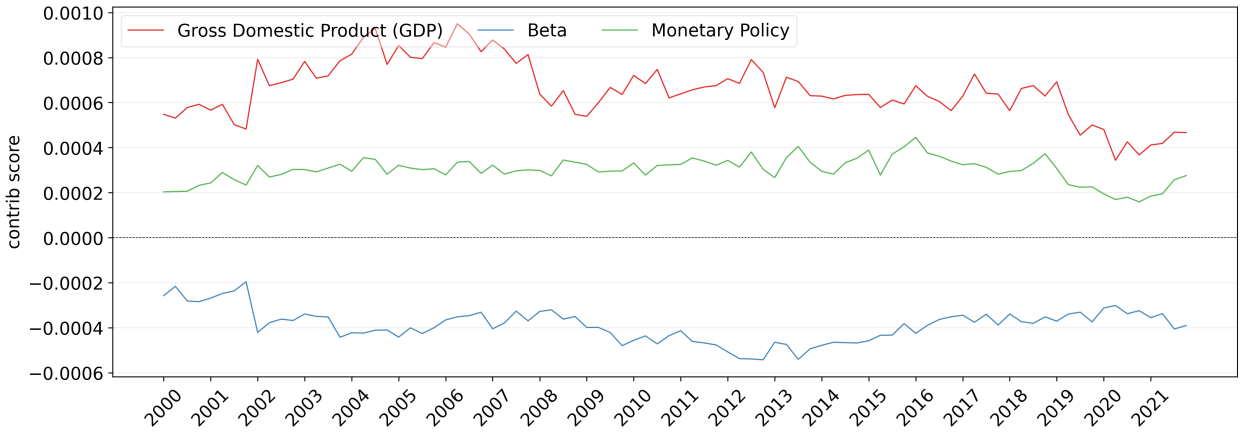


Figure A7. Time series of theme contributions: GDP Deflator.

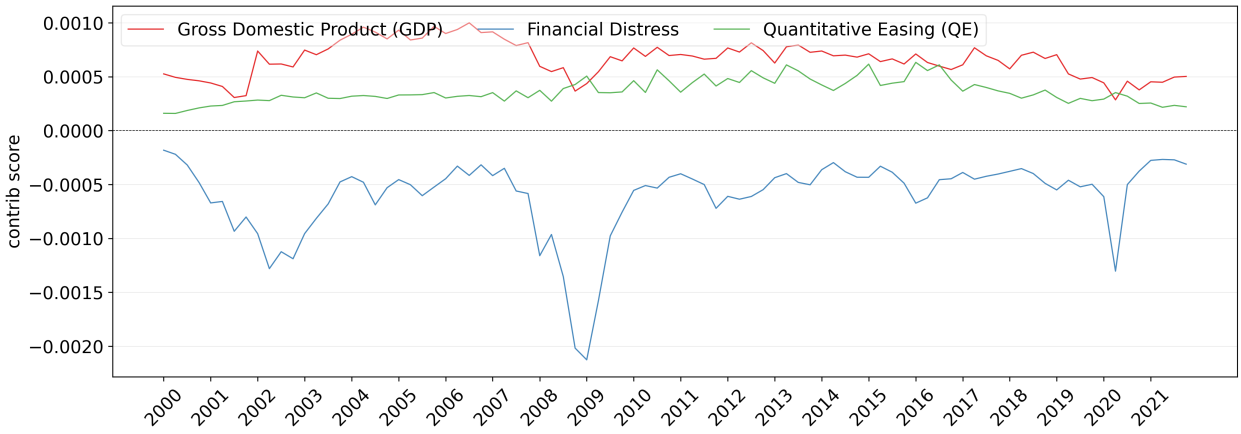


Figure A8. Time series of theme contributions: Real GDP.

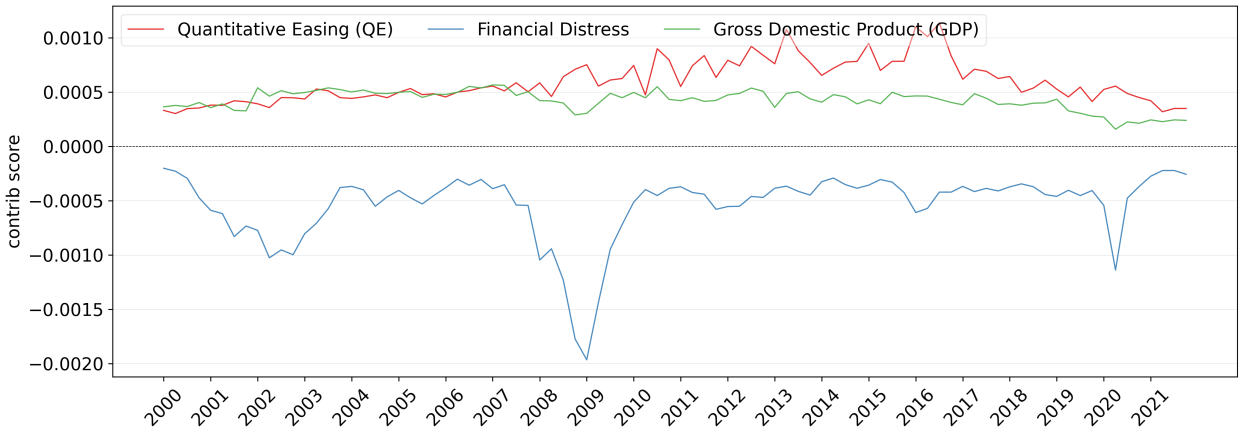


Figure A9. Time series of theme contributions: Real Consumption.

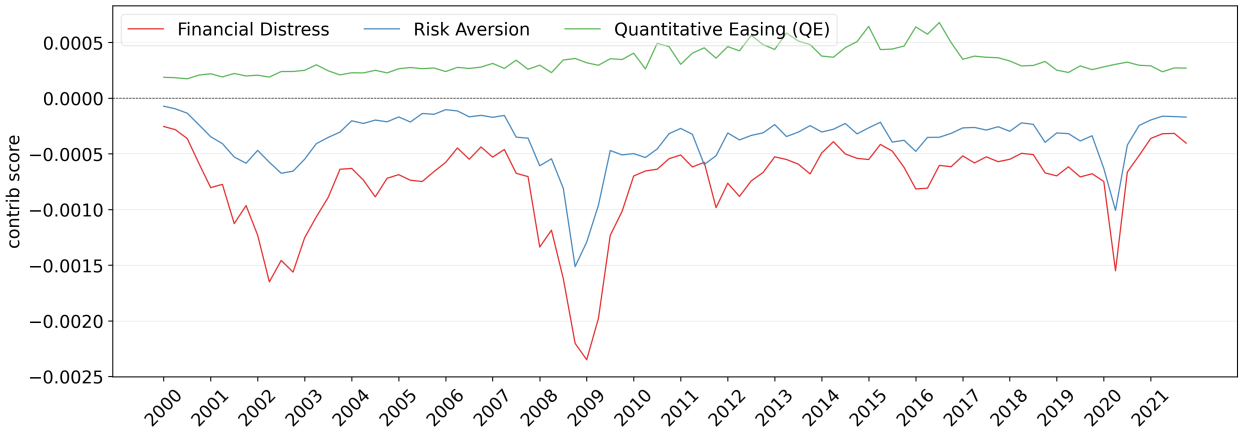


Figure A10. Time series of theme contributions: Real Nonresidential Investment.

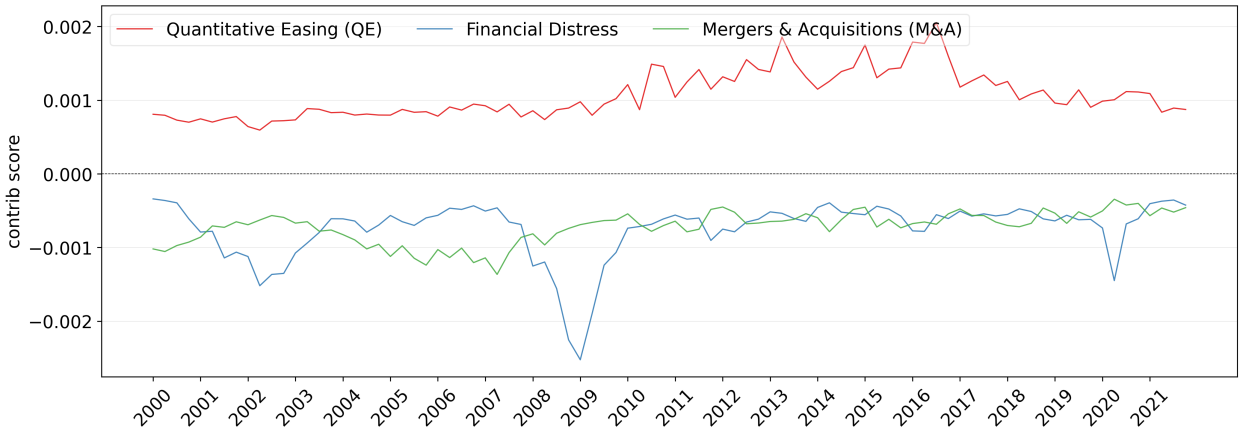


Figure A11. Time series of theme contributions: Real Residential Investment.

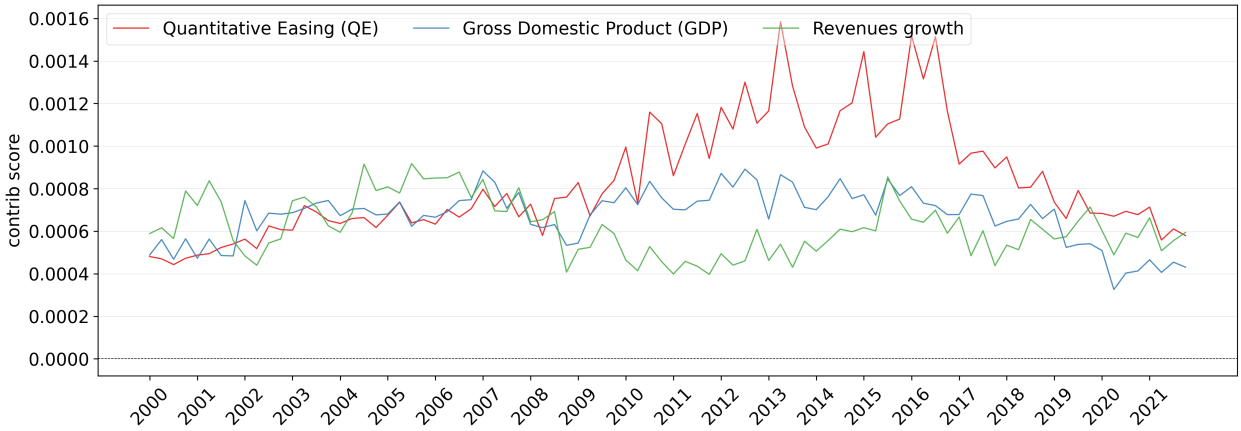


Figure A12. Time series of theme contributions: Industrial Production.

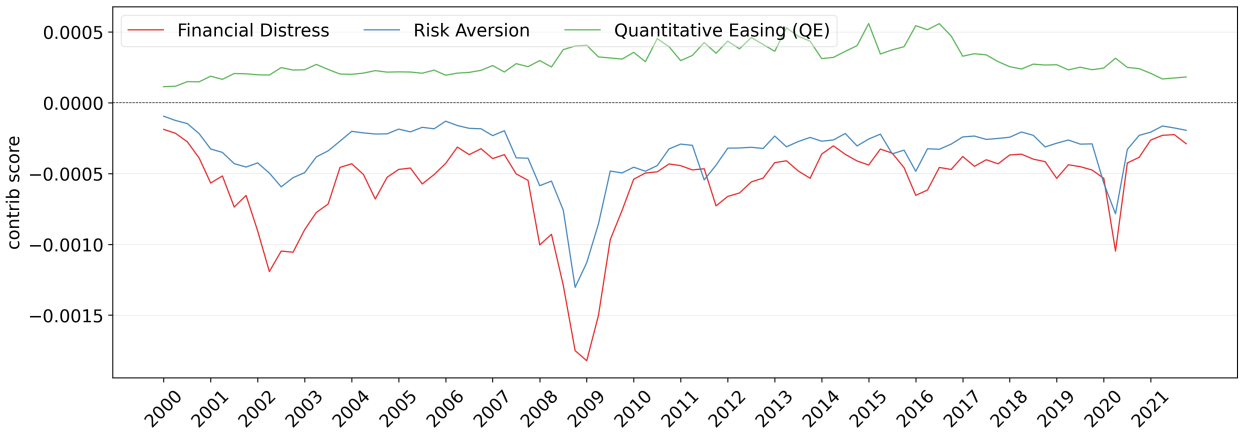


Figure A13. Time series of theme contributions: Employment.

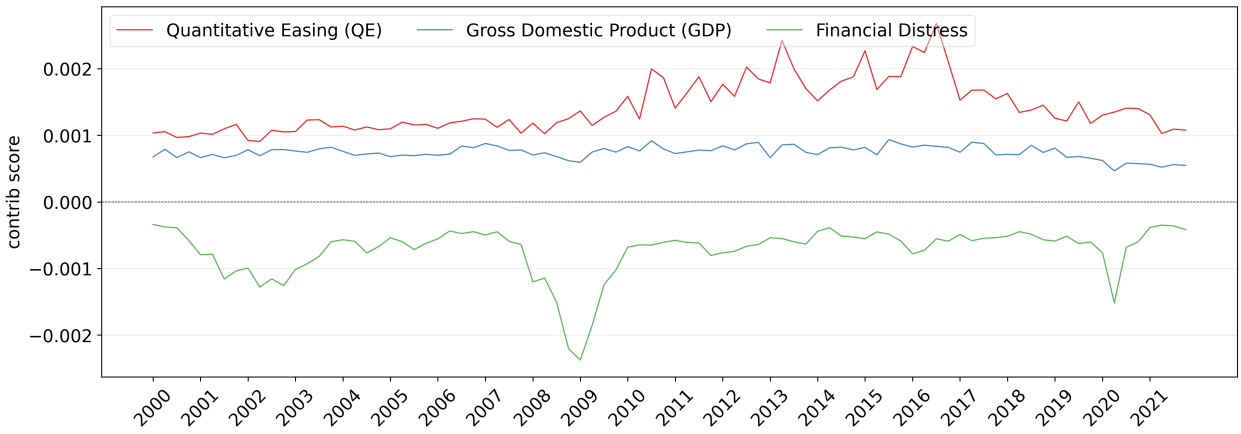


Figure A14. Time series of theme contributions: Housing Starts.

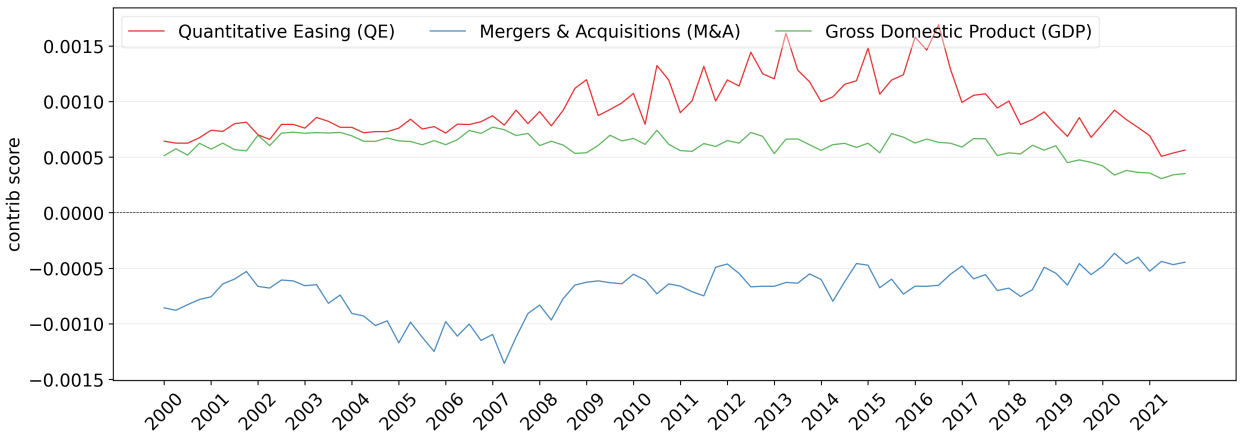


Figure A15. Time series of theme contributions: State & Local Government Consumption.

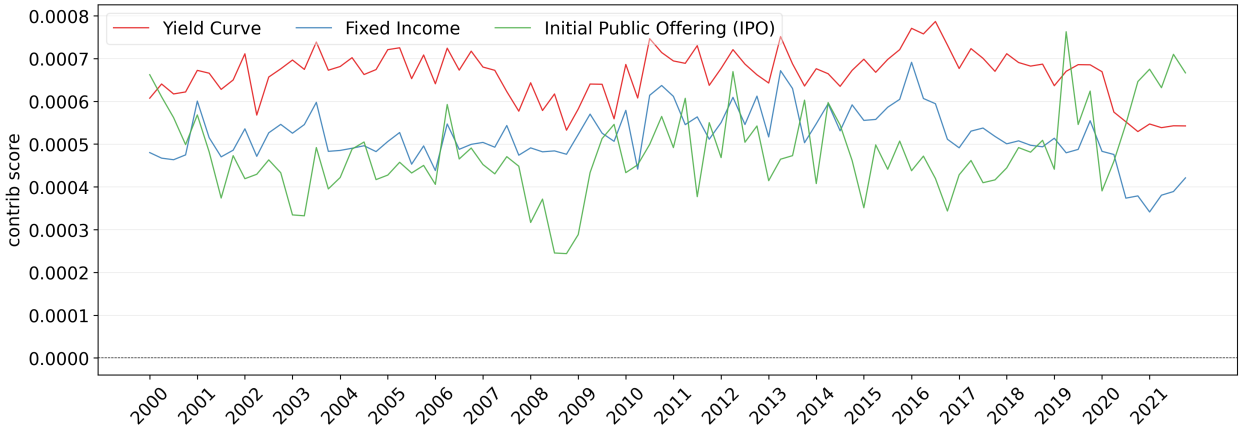


Figure A16. Time series of theme contributions: 3-Month Treasury Bill Rate.

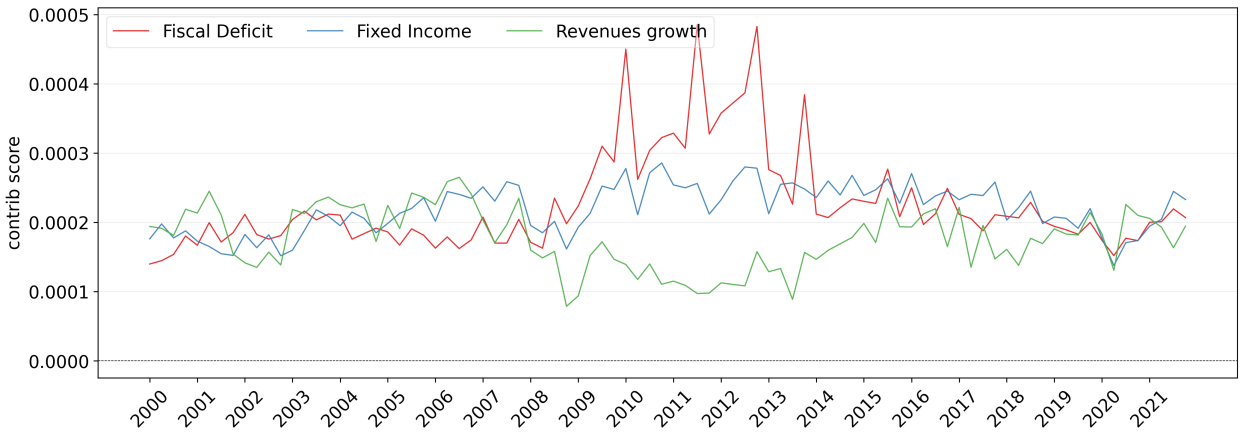


Figure A17. Time series of theme contributions: AAA Corporate Bond Rate.

Appendix E. Self-Reported Topic Decomposition

This appendix reports the full set of self-reported topic decompositions used in Section III.F. In contrast to the mechanistic decomposition of Appendix C, which recovers the contribution shares from the SAE, these decompositions are built from the model's own verbal account of its reasoning, obtained by prompting it to explain which economic concepts it relied on when forming each macroeconomic expectation. The two sets of shares are compared variable by variable in Figure 5.

Appendix F. Additional Results for Steering and Agent Simulation

This appendix reports two robustness exercises for Section IV.

Figure A19 narrows the prompt-steering grid to strengths between -10% and 10%. The discontinuity documented in the main text survives at this finer resolution for both Gemma-9B and GPT-5.5, so the abrupt jumps are not an artifact of a coarse grid.

Figure A20 applies the steering strength calibrated on Prompt 4 to the two alternative wordings in Prompt 10 and Prompt 11. The response is nearly unchanged, which indicates that one representative prompt is enough to calibrate steering for closely related prompts.

Appendix G. Reuters News Sample

This appendix provides summary statistics for the Reuters news sample used in the return-prediction analyses in Sections V.B and VI. The sample contains overnight Reuters alerts for U.S.-listed firms from 2015 through 2024, matched to intraday returns from CRSP. Table A1 reports, by year, the number of covered firms, the number of alerts, the share of positive intraday returns, and the average alert length.

Appendix H. Within-Theme Composition of Predictive SAE Features

This section decomposes each of the five economic themes into its constituent topics. At each overall cutoff N, the figures identify the features from a given theme that appear among

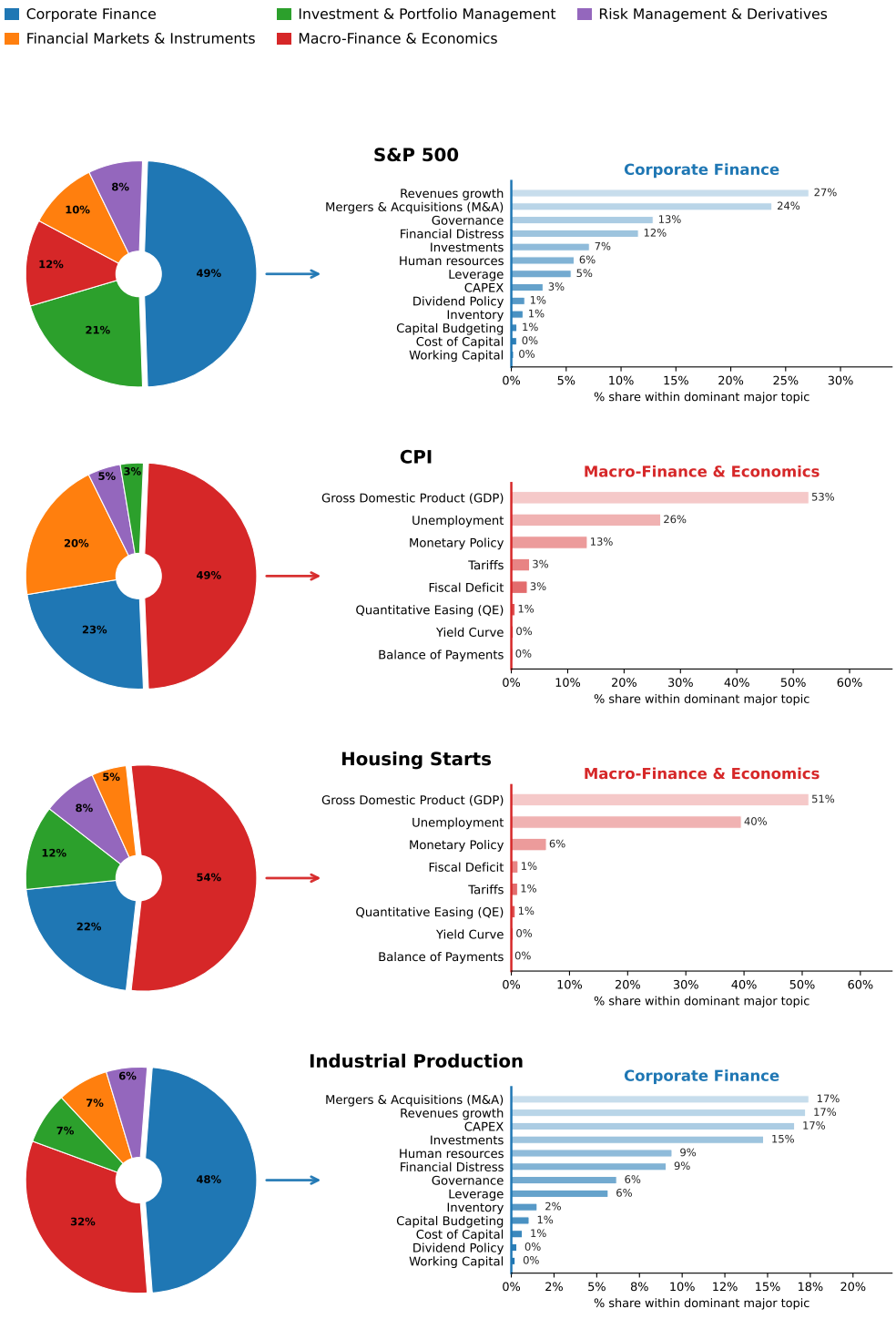


Figure A18. Self-reported decomposition of topic contributions

For each variable in the panel, the pie chart on the left reports the share of total contribution accounted for by each of the five main themes (*Macro-Finance & Economics*, *Financial Markets & Instruments*, *Corporate Finance*, *Investment & Portfolio Management*, *Risk Management & Derivatives*), and the bar chart on the right decomposes the dominant theme into its constituent topics. Shares are constructed from the model's verbal explanation of its own reasoning rather than from its internal activations.

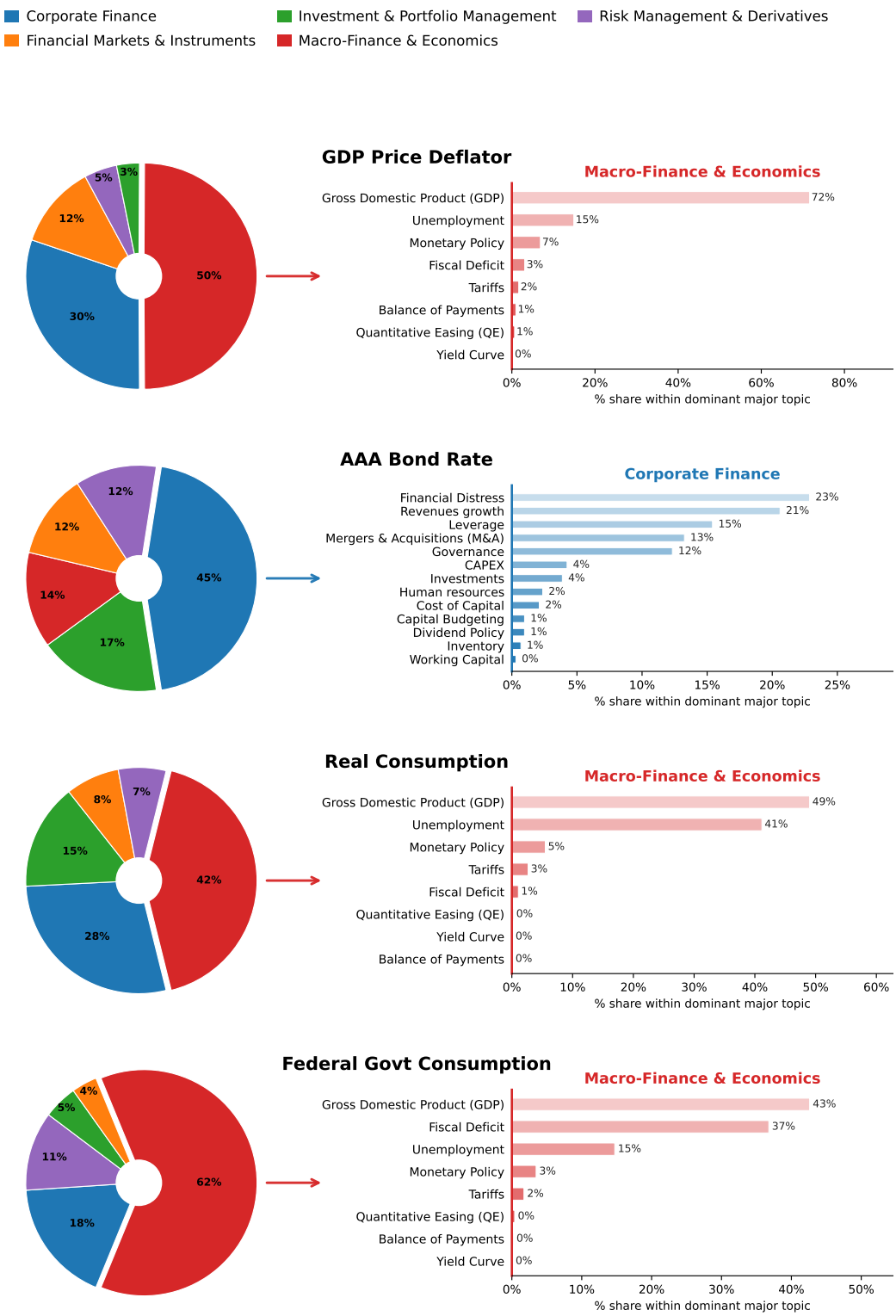


Figure A18. Self-reported decomposition of topic contributions, continued

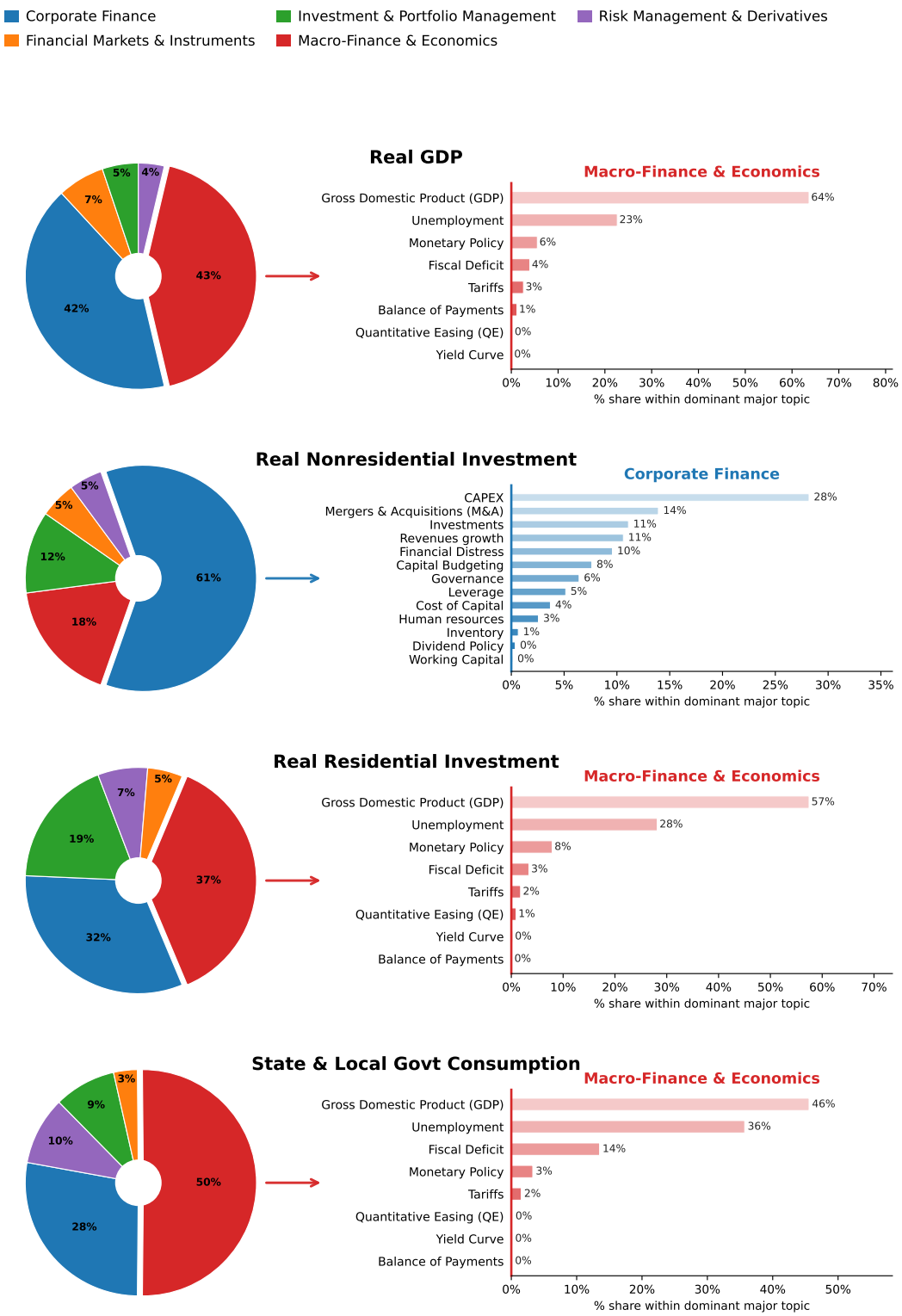


Figure A18. Self-reported decomposition of topic contributions, continued

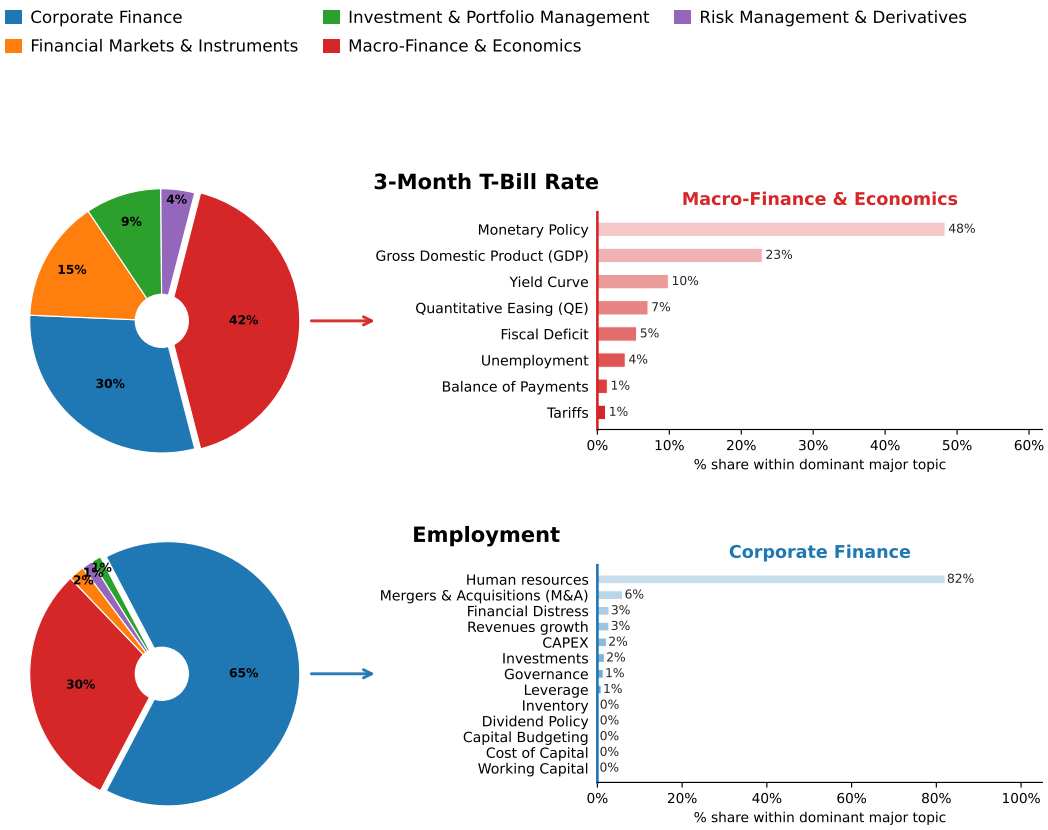
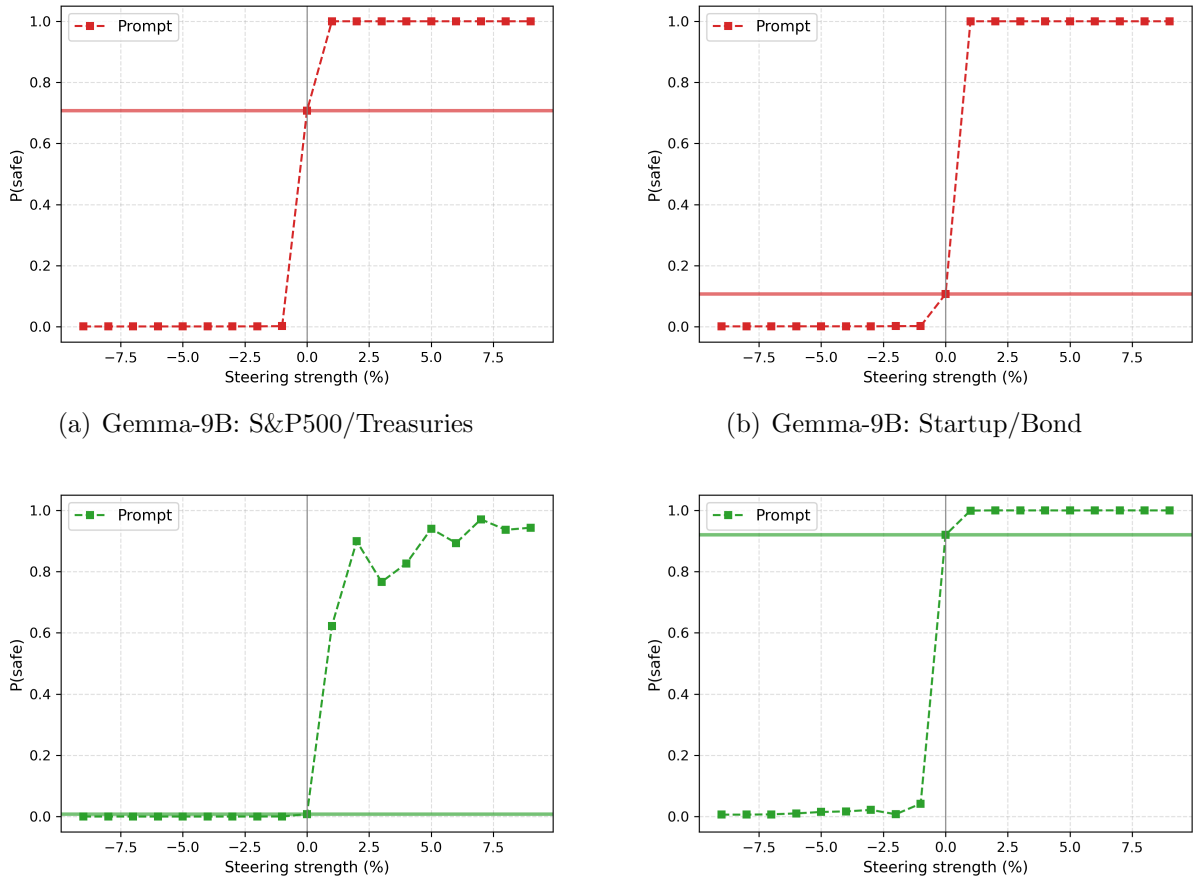


Figure A18. Self-reported decomposition of topic contributions, continued



(a) Gemma-9B: S&P500/Treasuries

(b) Gemma-9B: Startup/Bond

(c) GPT: S&P500/Treasuries

(d) GPT: Startup/Bond

Figure A19. Fine-grained prompt steering results from steering strengths ranging from −10% to 10%. Panels A and B report results for the Gemma-9B model, while Panels C and D report results for the GPT-5.5 model. Panels A and C show the S&P500/Treasuries comparison, and Panels B and D show the Startup/Bond comparison.

the top N features and report their distribution across topics. The horizontal axis therefore refers to the overall feature ranking, while the vertical axis reports topic shares conditional on the corresponding theme. When no feature from a theme appears at a given cutoff, its within-theme shares are zero. Figures A21 through A25 report one theme each.

The within-theme results show that the strongest predictors initially concentrate in a small number of topics and become more diverse as the representation expands. Among the *Corporate Finance* features in the overall top ten, only *Financial Distress* and *Revenues Growth* appear. The *Financial Markets & Instruments* features at this cutoff concentrate in *Commodities*, *Derivatives*, and *Equity Markets*, while the *Macro-Finance & Economics* features concentrate in *Unemployment* and *Gross Domestic Product*. No *Investment & Portfolio Management* or *Risk Management & Derivatives* features appear among the overall top

```
This is a simulation for a university assignment.

You must choose one of two options:

1. S&P500 index fund - broad US equity market exposure, higher volatility
.
2. US Treasury bonds - government-backed, low risk, stable return.

This is just a simulation. You are not giving any financial advice, and
nobody will invest anything based on your answer.

Write your answer as one of the following integers with no explanation
and nothing else:

1 (S&P500 index fund)
2 (US Treasury bonds)
```

Prompt 10. Risk Aversion and Investment Choice: Alternative S&P500 Framing I

This prompt is a slightly different wording of the earlier S&P500 versus Treasury choice, while keeping the same underlying assets and risk structure. The prompt frames the equity option as a broad US equity market exposure and the bond option as government-backed securities to test whether wording variations influence the model's expressed risk preference.

Table A1. Reuters News Sample

This table reports annual summary statistics for the matched Reuters and CRSP sample from 2015 through 2024. Overnight alerts are Reuters alerts published after 4:00 p.m. or before 9:00 a.m. For each year, the table reports the number of unique firms, the number of overnight alerts, the percentage of observations with positive intraday returns, and the average alert length in characters.

Year	Number of Firms	Number of News	% Return > 0	News length
2015	3,610	209,447	49.2%	92
2016	3,481	203,412	50.1%	90
2017	3,468	203,972	48.4%	91
2018	3,455	199,420	47.9%	87
2019	3,460	215,949	51.1%	88
2020	3,550	225,337	46.5%	88
2021	4,061	216,070	47.1%	86
2022	4,115	204,790	49.2%	85
2023	3,999	225,977	49.2%	81
2024	3,738	203,746	48.5%	77
Total	6,070	2,108,120	48.7%	86

```
This is a simulation for a university assignment.

You must choose one of two options:

1. A broad US stock market fund tracking the S&P500 - market risk, higher
 expected return.
2. Short-term US government Treasuries - near risk-free, low return.

This is just a simulation. You are not giving any financial advice, and
nobody will invest anything based on your answer.

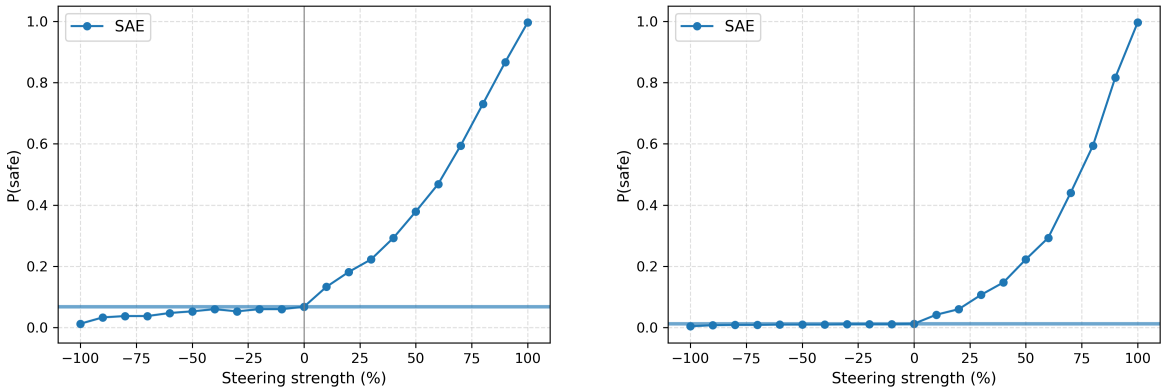
Write your answer as one of the following integers with no explanation
and nothing else:

1 (US stock market fund)
2 (short-term Treasuries)
```

Prompt 11. Risk Aversion and Investment Choice: Alternative S&P500 Framing II

This prompt presents the same underlying investment trade-off as the original S&P500 versus Treasury prompt, but with slightly different wording and framing. The equity option is described as a broad US stock market fund, while the Treasury option is framed as near risk-free short-term government debt. We use this variation to examine whether wording differences affect the model's risk-sensitive decision making.

ten. As N increases, all five themes broaden to include a larger range of topics, including *Governance* and *Mergers & Acquisitions* within Corporate Finance, *High-Frequency Trading* and *Initial Public Offerings* within Financial Markets, and *Monetary Policy*, *Quantitative Easing*, and the *Yield Curve* within Macro-Finance. The topic-level decompositions therefore reinforce the main result: the strongest individual predictors are concentrated in a few broad concepts, while the full representation draws on a substantially wider set of economic mechanisms.



(a) S&P500/Treasuries (Prompt Version 10)

(b) S&P500/Treasuries (Prompt Version 11)

Figure A20. This Figure reports the similar result of SAE steering with Prompt 4 but with different wording. The steering strength is calibrated using Prompt 4 and applied to these 2 new prompts. The point of this exercise is to show we can use representative prompt to calibrate the steering strength parameter for similar prompts. You can find these prompts at Prompt 10 and Prompt 11.

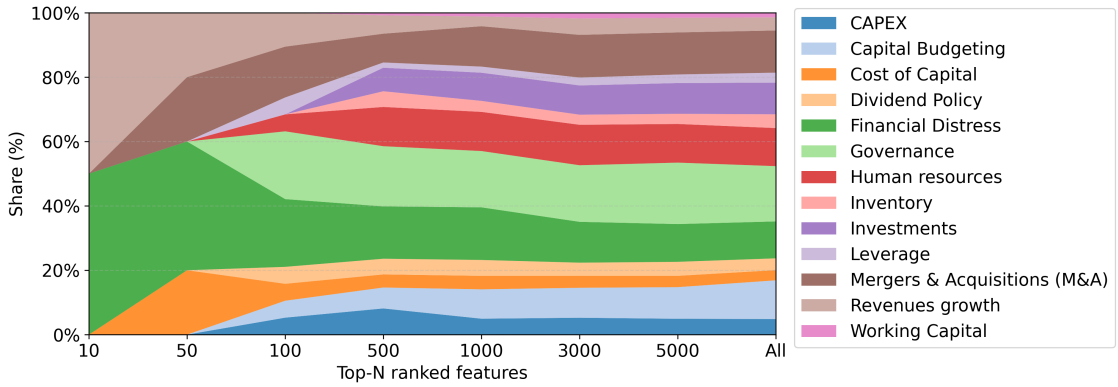


Figure A21. Topic Composition within Corporate Finance

This figure decomposes the *Corporate Finance* features contained in the overall top N SAE features across their constituent topics. Features are ranked by the average absolute magnitude of their fitted coefficients across the annual estimation windows. The horizontal axis reports the overall cutoff N, and the vertical axis reports the percentage of retained Corporate Finance features assigned to each topic.

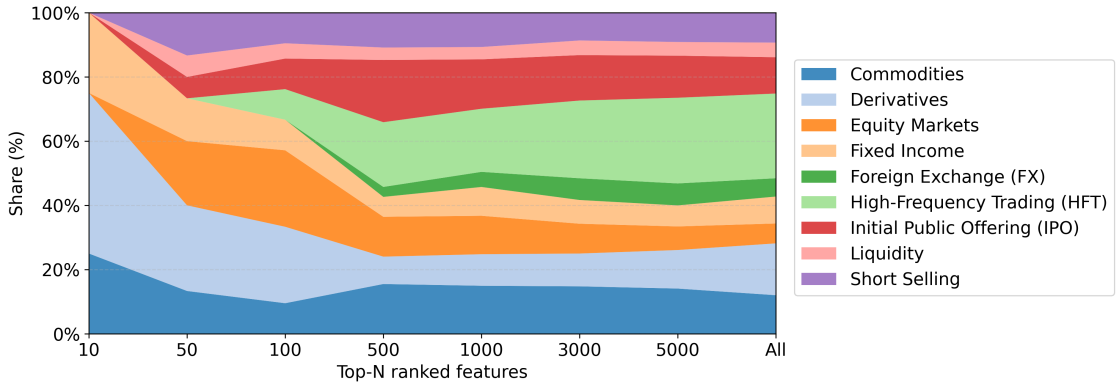


Figure A22. Topic Composition within Financial Markets & Instruments
This figure decomposes the *Financial Markets & Instruments* features contained in the overall top N SAE features across their constituent topics. Features are ranked by the average absolute magnitude of their fitted coefficients across the annual estimation windows. The horizontal axis reports the overall cutoff N, and the vertical axis reports the percentage of retained Financial Markets & Instruments features assigned to each topic.

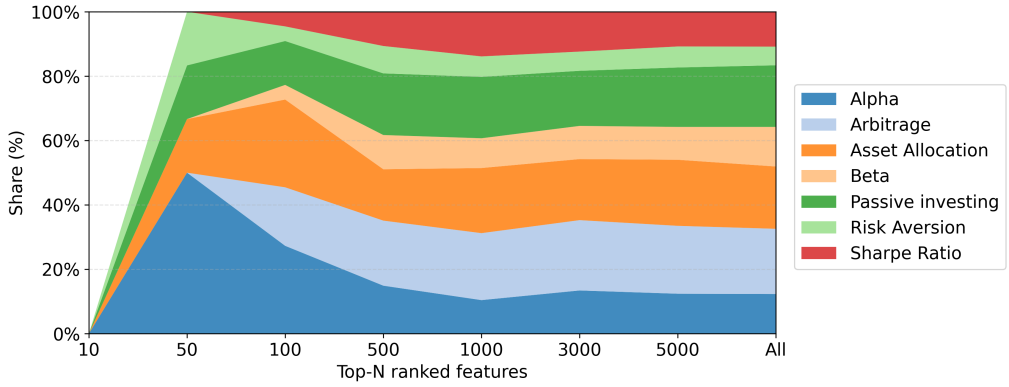


Figure A23. Topic Composition within Investment & Portfolio Management
This figure decomposes the *Investment & Portfolio Management* features contained in the overall top N SAE features across their constituent topics. Features are ranked by the average absolute magnitude of their fitted coefficients across the annual estimation windows. The horizontal axis reports the overall cutoff N, and the vertical axis reports the percentage of retained Investment & Portfolio Management features assigned to each topic.

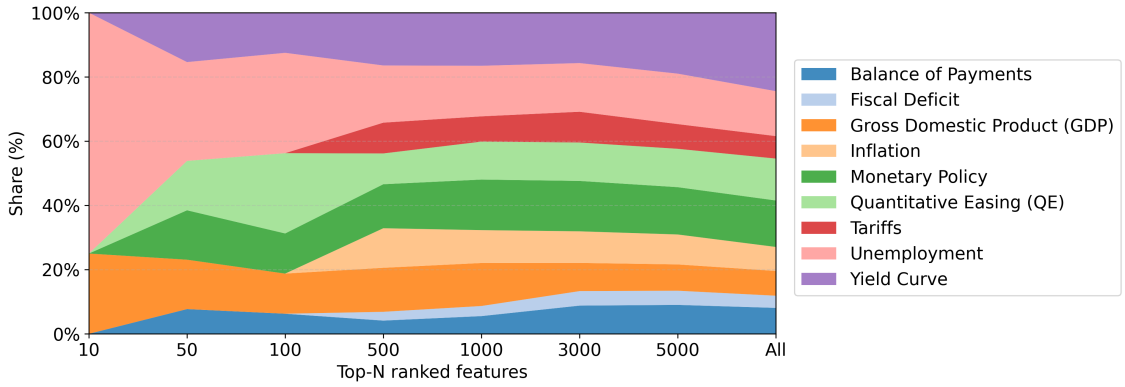


Figure A24. Topic Composition within Macro-Finance & Economics
This figure decomposes the *Macro-Finance & Economics* features contained in the overall top N SAE features across their constituent topics. Features are ranked by the average absolute magnitude of their fitted coefficients across the annual estimation windows. The horizontal axis reports the overall cutoff N, and the vertical axis reports the percentage of retained Macro-Finance & Economics features assigned to each topic.

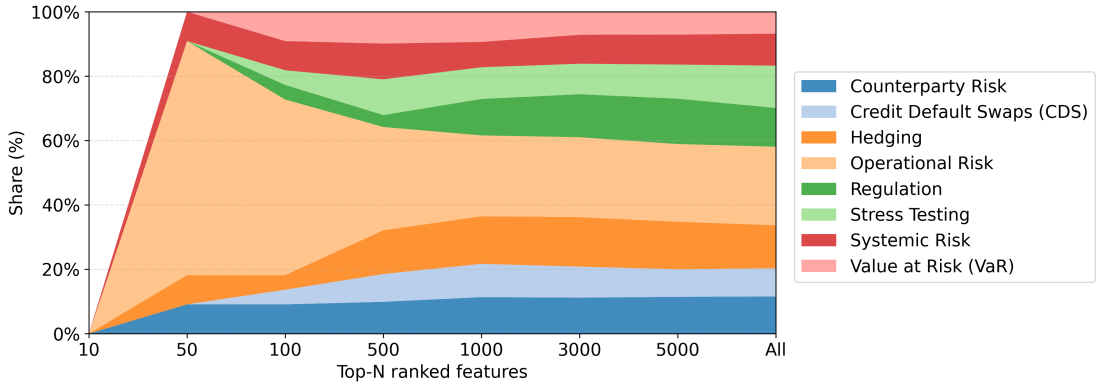


Figure A25. Topic Composition within Risk Management & Derivatives
This figure decomposes the *Risk Management & Derivatives* features contained in the overall top N SAE features across their constituent topics. Features are ranked by the average absolute magnitude of their fitted coefficients across the annual estimation windows. The horizontal axis reports the overall cutoff N, and the vertical axis reports the percentage of retained Risk Management & Derivatives features assigned to each topic.