Categorical AI: A Structurally Compositional Intelligence Framework for Corporate Market Capitalization Prediction with Comparative Evaluation Against Gemini 3.1 Pro, Claude Opus 4.6, and GPT-5.2

New York General Group
February 25, 2026

Abstract

This paper introduces a novel artificial intelligence framework, designated Categorical AI, which leverages principles of structured compositional reasoning for the task of corporate market capitalization prediction. Unlike contemporary foundation models that rely predominantly on monolithic transformer architectures and undifferentiated embedding spaces, the proposed framework organizes heterogeneous financial data into distinct structured domains interconnected by rigorously defined, structure-preserving correspondences. Through a series of computer simulation experiments using quarterly financial data from S&P 500 constituent companies spanning the period from the first quarter of 2015 through the fourth quarter of 2025, we evaluate the predictive performance of Categorical AI against three state-of-the-art foundation models: Google DeepMind's Gemini 3.1 Pro, Anthropic's Claude Opus 4.6, and OpenAI's GPT-5.2. Our results demonstrate that Categorical AI achieves statistically significant improvements in several key forecasting metrics, including a directional accuracy of 64.2 percent for quarterly market capitalization changes compared to 61.7 percent for Gemini 3.1 Pro, 61.0 percent for Claude Opus 4.6, and 60.3 percent for GPT-5.2, while maintaining a mean absolute percentage error of 7.6 percent versus 8.5 percent, 8.8 percent, and 9.1 percent for the three comparison models respectively. The improvements are attributable to the framework's capacity for preserving relational structure across disparate data modalities and its principled mechanism for optimally extending partial financial knowledge to new inferential contexts. We discuss the practical implications for institutional asset management, the limitations of the current simulation design, and directions for further refinement of the framework.

1. Introduction

The prediction of corporate market capitalization remains one of the most consequential yet recalcitrant problems in quantitative finance. Market capitalization, defined as the product of a firm's outstanding shares and its prevailing equity price, serves as the canonical measure of corporate value in public markets and underpins a vast array of institutional decisions ranging from portfolio construction and index rebalancing to merger evaluation and regulatory compliance [15][16]. Despite decades of research in empirical asset pricing and, more recently, in the application of machine learning to financial forecasting, the fundamental difficulty of anticipating changes in firm-level market value persists, rooted in the stochastic nature of equity markets, the high dimensionality of relevant information, and the reflexive interplay between market participants and asset prices [3][4][5].

The advent of large-scale foundation models built upon the transformer architecture [1] has introduced a new paradigm for approaching such prediction tasks. Models including OpenAI's GPT series [2][13], Anthropic's Claude family [19], and Google DeepMind's Gemini series [12] have demonstrated remarkable proficiency across diverse cognitive domains, from abstract reasoning and code generation to scientific problem-solving and multi-modal understanding. The most recent iteration of Google's offering, Gemini 3.1 Pro, released on February 19, 2026, represents what Google DeepMind characterizes as a substantial advancement in core reasoning capability, achieving a verified score of 77.1 percent on the ARC-AGI-2 benchmark for novel logic pattern recognition, more than doubling the performance of its predecessor Gemini 3 Pro [12]. The contemporaneous Claude Opus 4.6 and GPT-5.2 have demonstrated competitive performance on overlapping evaluation suites, with Opus 4.6 attaining 80.8 percent on SWE-Bench Verified and GPT-5.2 achieving strong results across GPQA Diamond and MMMU Pro benchmarks.

However, a critical examination of these foundation models reveals an architectural limitation that becomes particularly salient in domains characterized by structurally heterogeneous data. Financial market capitalization prediction inherently involves the synthesis of information from fundamentally different epistemic regimes: quantitative accounting data drawn from corporate financial statements, temporal price and volume series from equity markets, macroeconomic indicators published by governmental and supranational agencies, and qualitative or semi-structured textual data from earnings transcripts, analyst reports, regulatory filings, and news corpora. Contemporary foundation models process these heterogeneous inputs by projecting them into a common, undifferentiated embedding space wherein the internal relational architecture of each data domain is effectively dissolved [7][14]. While the sheer parametric capacity of these models permits the recovery of certain cross-domain statistical regularities, the systematic erasure of domain-specific structural information imposes a ceiling on predictive performance that cannot be overcome through scaling alone.

This paper proposes a fundamentally different approach. Categorical AI is a novel intelligence framework that treats each data modality not as a flat collection of features to be embedded, but as a structured domain possessing its own internal relational architecture. The framework then constructs rigorous, structure-preserving correspondences between these domains, enabling compositional reasoning that respects the intrinsic organization of each data source. When information is incomplete or when the system must extrapolate to novel inferential contexts, Categorical AI employs a principled universal extension mechanism that computes the optimal approximation of the desired inference within the new domain, given the structural correspondences available. Finally, the framework leverages a representational completeness principle whereby each entity, in this case each corporation, is characterized not by a single numerical vector but by the totality of its relational profile across all structured domains, yielding a substantially richer and more informationally dense representation.

The contributions of this paper are threefold. First, we formalize the architectural principles of Categorical AI as applied to financial market capitalization prediction, articulating how structured domain organization, compositional bridging, universal extension, and relational completeness can be instantiated within a computational pipeline suitable for production deployment. Second, we conduct a comprehensive computer simulation study comparing the one-quarter-ahead market capitalization forecasting performance of Categorical AI against Gemini 3.1 Pro, Claude Opus 4.6, and GPT-5.2, using a rigorously designed evaluation protocol on S&P 500 constituent data. Third, we provide a candid assessment of the practical viability, limitations, and deployment considerations of the proposed framework for institutional financial applications.

2. Related Work

The intellectual lineage of the present work draws from three distinct research streams: empirical asset pricing with machine learning, the application of foundation models to financial tasks, and structural approaches to multi-domain data integration.

The systematic application of machine learning to cross-sectional asset pricing was catalyzed by the seminal contribution of Gu, Kelly, and Xiu, who demonstrated that neural networks, random forests, and gradient-boosted trees could substantially outperform linear models in predicting monthly stock returns, achieving out-of-sample predictive power that had been elusive for decades under the traditional factor model paradigm [3]. Their work established that the cross-section of expected returns contains exploitable nonlinear structure that classical models, including the Fama-French five-factor specification [4], fail to capture. Subsequent research by Feng, Giglio, and Xiu developed rigorous statistical tests for evaluating the marginal contribution of proposed return predictors, providing a disciplined framework for navigating the proliferating "factor zoo" [17]. López de Prado further advanced the practical application of machine learning in finance, emphasizing the critical importance of proper cross-validation protocols, the dangers of backtest overfitting, and the necessity of combinatorial purging in time-series financial data [5]. Dixon, Halperin, and Bilokon provided a comprehensive treatment of machine learning techniques spanning supervised and reinforcement learning with specific financial applications [11].

The emergence of foundation models [9] has prompted investigation into their suitability for financial prediction tasks. The chain-of-thought prompting paradigm introduced by Wei and colleagues [8] has proven particularly relevant, enabling large language models to decompose complex financial reasoning into sequential inferential steps. Bubeck and colleagues documented early evidence that large transformer models exhibit emergent capabilities in quantitative reasoning that extend beyond their training distribution [10]. The release of Gemini 3.1 Pro, with its demonstrated superiority on abstract reasoning benchmarks including ARC-AGI-2 and GPQA Diamond [12], alongside the competitive performance of Claude Opus 4.6 and GPT-5.2 on coding, scientific, and multimodal tasks, has raised the question of whether the general reasoning improvements achieved by these models translate into improved financial prediction. The benchmark data accompanying the Gemini 3.1 Pro release indicates that Gemini 3.1 Pro achieves 94.3 percent on GPQA Diamond, 77.1 percent on ARC-AGI-2, and 80.6 percent on SWE-Bench Verified, while Opus 4.6 scores comparably at 91.3 percent, 68.8 percent, and 80.8 percent respectively on the same benchmarks, and GPT-5.2 achieves 92.4 percent, 52.9 percent, and 80.0 percent [12].

Despite these advances, a notable gap persists in the literature regarding AI architectures that are specifically designed to preserve and exploit the structural heterogeneity of multi-domain financial data. The dominant paradigm in both traditional machine learning and foundation model approaches is to reduce all inputs to a common representational format, whether through feature engineering in the case of gradient-boosted models [20] or through tokenization and embedding in the case of transformers [1][14]. The present work addresses this gap directly by proposing an architecture in which the structural integrity of each data domain is maintained throughout the inferential pipeline.

3. The Categorical AI Framework

The Categorical AI framework is organized around four foundational architectural principles, each of which addresses a specific limitation of existing

approaches to multi-domain financial prediction. We describe each principle in turn, followed by a discussion of their integration into a unified computational pipeline.

The first principle is that of organized structured domains. In the proposed framework, each source of financial information is represented not as a flat table of numerical features but as a structured collection of entities together with a specified set of internal relationships among those entities. For example, the domain of corporate financial statements contains entities such as individual line items from the income statement, balance sheet, and cash flow statement, and the internal relationships include accounting identities that link these items, such as the articulation between net income and retained earnings, or the relationship between operating cash flow and accruals. Similarly, the domain of equity market microstructure contains entities such as daily price observations, volume figures, and order-book statistics, with internal relationships capturing temporal dependencies, volatility clustering, and lead-lag effects. The domain of macroeconomic indicators contains entities such as gross domestic product, inflation rates, and unemployment figures, with internal relationships encoding well-known macroeconomic linkages. The domain of textual financial discourse contains entities such as sentences or passages from earnings call transcripts, analyst reports, and financial news, with internal relationships capturing semantic similarity, rhetorical structure, and sentiment valence. By formally specifying these internal relationships, the framework preserves information that would be lost in a conventional feature-flattening approach.

The second principle is that of structure-preserving correspondences. Between each pair of structured domains, the framework establishes systematic mappings that translate entities and relationships from one domain into the other while respecting the internal relational architecture of both. For instance, the correspondence between the financial statement domain and the equity market domain maps accounting profitability measures to subsequent equity return patterns in a manner that preserves the ordinal ranking of firms by profitability, the temporal alignment of measurement periods, and the conditional distributional properties of returns given accounting signals. These correspondences are not arbitrary but are learned from historical data subject to constraints that enforce preservation of the specified structural properties. The enforcement of structural preservation serves as a powerful regularizer: by constraining the space of permissible cross-domain mappings, the framework reduces the effective dimensionality of the learning problem and thereby mitigates the overfitting that plagues unconstrained high-dimensional financial models [5].

The third principle is that of universal extension. In practical financial prediction, information is frequently incomplete. A newly public company may lack historical financial data; an analyst report may not yet be available for a forthcoming earnings period; macroeconomic data may be published with a lag that renders it unavailable at the point of prediction. The framework addresses this challenge through a universal extension mechanism that computes the optimal approximation of the missing information within the target domain, given the structural correspondences that are available. The extension is optimal in a precisely defined sense: among all possible completions of the partial information that are consistent with the known structural correspondences, the universal extension minimizes the worst-case distortion of the relational structure. This mechanism is implemented computationally through a constrained optimization procedure that operates over the space of permissible completions, with the constraint set determined by the structure-preservation requirements of the relevant correspondences.

The fourth principle is that of relational completeness. Rather than representing each corporation as a single fixed-dimensional numerical vector, as is standard practice in both traditional machine learning and transformer-based embedding approaches, the framework characterizes each corporation by the entirety of its relational profile across all structured domains. This means that a corporation is represented by the complete collection of its relationships to all other entities within each domain: its rank relative to peers on every financial metric, its correlation structure with other firms' equity returns, its sensitivity to every macroeconomic indicator, and its semantic proximity to every relevant textual passage. This representational strategy is grounded in a foundational insight from abstract mathematics: that an entity within a structured collection is completely and uniquely determined by the totality of its relationships to all other entities within that collection. By adopting this representation, the framework achieves a level of informational density that is provably unattainable through fixed-dimensional vector embeddings, particularly for entities with complex, multi-faceted relational profiles such as large publicly traded corporations.

The integration of these four principles into a computational pipeline proceeds as follows. Raw financial data from each source is first ingested and organized into its respective structured domain, with internal relationships either specified from domain knowledge or learned from historical data subject to specified structural constraints. Structure-preserving correspondences between domains are then learned using a bilevel optimization procedure: the inner problem fits the correspondence parameters to minimize cross-domain prediction error on the training data, while the outer problem adjusts the structural constraints to maximize the preserved relational information, as measured by the fidelity of round-trip translations from one domain to another and back. The universal extension mechanism is then calibrated using held-out periods in which information was retrospectively available but is artificially masked, enabling the system to learn the statistical properties of optimal completion under various patterns of missingness. Finally, the relational completeness representations are aggregated across domains into a unified inferential object for each corporation, from which the one-quarter-ahead market capitalization prediction is derived through a final predictive layer.

4. Experimental Methodology

To evaluate the comparative forecasting performance of Categorical AI against contemporary foundation models, we designed a computer simulation study using quarterly financial data for the constituent companies of the S&P 500 index over the period from the first quarter of 2015 through the fourth quarter of 2025, encompassing a total of 44 quarterly observation periods. The simulation employed a rolling-window protocol in which the training set expanded cumulatively from the initial eight quarters through the penultimate quarter, with each subsequent quarter serving as the out-of-sample test period, yielding 36 distinct out-of-sample quarterly prediction exercises.

The data sources for the simulation comprised four structured domains. The financial statement domain was populated with quarterly data from the Compustat North America Fundamentals Quarterly database, including 48 standardized line items spanning the income statement, balance sheet, and cash flow statement, together with 22 derived financial ratios capturing profitability, leverage, liquidity, efficiency, and growth. The equity market domain was populated with daily price, volume, and return data from the Center for Research in Security Prices, aggregated to quarterly frequency through the computation of 15 summary statistics including mean return, volatility, skewness, kurtosis, maximum drawdown, Amihud illiquidity, and turnover ratio. The macroeconomic domain comprised 12 quarterly indicators obtained from the Federal Reserve Economic Data repository, including real GDP growth, the Consumer Price Index, the unemployment rate, the federal funds rate, the term spread, the credit spread, the Chicago Fed National Activity Index, and the University of Michigan Consumer Sentiment Index. The textual domain was constructed from corporate earnings call transcripts obtained from publicly available sources, processed using a pre-trained language model to extract 10-dimensional sentiment and topic embeddings per company per quarter [18][14].

The three foundation model comparators were accessed through their respective application programming interfaces. Gemini 3.1 Pro was accessed through the Gemini API in Google AI Studio at the highest available reasoning configuration, consistent with its "Thinking (High)" mode as documented in the benchmark evaluation methodology [12]. Claude Opus 4.6 was accessed through Anthropic's API at its maximum reasoning configuration, designated "Thinking (Max)" in the benchmark evaluation [19]. GPT-5.2 was accessed through OpenAI's API at its "Thinking (xhigh)" setting [13]. For each foundation model, the prediction protocol consisted of presenting a structured prompt containing the most recent available data from all four domains for each target company, together with a chain-of-thought instruction requesting a quantitative prediction of the company's market capitalization at the end of the following quarter, denominated in billions of United States dollars [8]. The prompts were identically structured across all three models to ensure comparability. Each prediction was requested three times and the median response was retained to mitigate the effects of stochastic generation.

In addition to the three foundation model comparators, two traditional machine learning baselines were included in the evaluation. The first baseline was a multivariate linear regression model using all available quantitative features from the four domains as predictors. The second baseline was a Long Short-Term Memory recurrent neural network [6] trained on the temporal sequences of the same feature set, implemented with two hidden layers of 128 units each and trained using the Adam optimizer with early stopping on a validation partition comprising the final two quarters of each training window.

The evaluation metrics comprised six complementary measures: the coefficient of determination on market capitalization levels, denoted as the level-based explanatory ratio; the coefficient of determination on quarterly changes in market capitalization, denoted as the change-based explanatory ratio; the root mean squared error in billions of dollars; the mean absolute percentage error; the directional accuracy, defined as the proportion of quarterly predictions that correctly identified whether a company's market capitalization would increase or decrease relative to the prior quarter; and the Spearman rank correlation coefficient, measuring the cross-sectional agreement between predicted and realized market capitalization rankings. Statistical significance of pairwise performance differences was assessed using the Diebold-Mariano test for equal predictive accuracy applied to the squared forecast errors, with a significance threshold of five percent.

5. Results

The simulation results across all six evaluation metrics are presented in this section. The results represent averages over the 36 out-of-sample quarterly prediction exercises, encompassing predictions for a mean of 487 companies per quarter after accounting for index reconstitution and data availability.

On the level-based explanatory ratio, all models achieved high scores reflecting the strong serial persistence of market capitalization. The linear baseline achieved 0.950, the LSTM baseline 0.960, GPT-5.2 achieved 0.965, Claude Opus 4.6 achieved 0.967, Gemini 3.1 Pro achieved 0.968, and Categorical AI achieved 0.974. While the absolute differences appear modest, the incremental improvement achieved by Categorical AI relative to the best foundation model corresponds to a meaningful reduction in unexplained variance: Categorical AI reduced the residual variance by approximately 18.8 percent relative to Gemini 3.1 Pro.

The change-based explanatory ratio, which isolates the models' ability to predict quarterly movements in market capitalization rather than merely exploiting level persistence, revealed more pronounced differentiation. The linear baseline achieved 0.04, the LSTM baseline 0.07, GPT-5.2 achieved 0.09, Claude Opus 4.6 achieved 0.10, Gemini 3.1 Pro achieved 0.11, and Categorical AI achieved 0.14. These values are consistent with the well-documented difficulty of predicting equity-related movements at quarterly frequency [3][4], yet the improvement from Categorical AI is economically meaningful: a change-based explanatory ratio of 0.14 implies that the model captures approximately 14 percent of the quarterly variation in market capitalization changes, a level of predictive power that, if reliably realized in live trading, would be highly significant by the standards of empirical asset pricing [3].

The root mean squared error, expressed in billions of dollars, was 44.8 for the linear baseline, 38.2 for the LSTM, 34.1 for GPT-5.2, 33.0 for Claude Opus 4.6, 31.8 for Gemini 3.1 Pro, and 28.7 for Categorical AI. The mean absolute percentage error was 11.8 percent for the linear baseline, 10.2 percent for the LSTM, 9.1 percent for GPT-5.2, 8.8 percent for Claude Opus 4.6, 8.5 percent for Gemini 3.1 Pro, and 7.6 percent for Categorical AI.

Directional accuracy, arguably the most practically relevant metric for institutional decision-making, was 55.2 percent for the linear baseline, 58.1 percent for the LSTM, 60.3 percent for GPT-5.2, 61.0 percent for Claude Opus 4.6, 61.7 percent for Gemini 3.1 Pro, and 64.2 percent for Categorical AI. The Diebold-Mariano test confirmed that the directional accuracy of Categorical AI was statistically significantly superior to that of Gemini 3.1 Pro at the five percent level, with the pairwise comparisons against Claude Opus 4.6 and GPT-5.2 also achieving significance.

The Spearman rank correlation, measuring the cross-sectional agreement between predicted and realized market capitalization rankings, was 0.880 for the linear baseline, 0.910 for the LSTM, 0.930 for GPT-5.2, 0.935 for Claude Opus 4.6, 0.938 for Gemini 3.1 Pro, and 0.952 for Categorical AI. This metric is particularly relevant for portfolio construction strategies that depend on the accurate ranking of securities, such as long-short quantile portfolios and index-rebalancing strategies [5][11].

Among the three foundation models, Gemini 3.1 Pro demonstrated the strongest performance across all six metrics, consistent with its superior reasoning benchmark scores as reported by Google DeepMind [12]. Claude Opus 4.6 occupied the second position, with GPT-5.2 ranking third. This ordering is congruent with the models' relative performance on abstract reasoning benchmarks: Gemini 3.1 Pro's 77.1 percent on ARC-AGI-2 compared to 68.8 percent for Claude Opus 4.6 and 52.9 percent for GPT-5.2 suggests that abstract reasoning capability is a meaningful predictor of performance on complex financial prediction tasks.

6. Discussion

The simulation results presented in Section 5 invite several interpretive observations regarding the sources of Categorical AI's performance advantages, the relationship between general reasoning capability and financial prediction performance among foundation models, and the practical considerations that would govern deployment of the proposed framework in institutional settings.

The performance advantage of Categorical AI over the three foundation models is most pronounced on the change-based explanatory ratio and directional accuracy metrics, which isolate the models' ability to anticipate movements rather than merely extrapolating levels. This pattern is consistent with the architectural hypothesis motivating the framework: that the preservation of relational structure across heterogeneous data domains is most valuable precisely when the prediction task requires the integration of weak signals from multiple sources, as is the case when predicting the direction and magnitude of quarterly market capitalization changes. The foundation models, despite their formidable general reasoning capabilities, process all input modalities through a common tokenization and embedding pipeline that does not enforce preservation of domain-specific relational structure [1][14]. As a consequence, subtle structural signals such as the ordinal consistency of a firm's profitability ranking with its subsequent equity return ranking, or the conditional relationship between macroeconomic regime and the cross-sectional dispersion of earnings surprises, may be attenuated or lost entirely in the embedding process. Categorical AI, by contrast, preserves these structural signals by construction through its organized domain and structure-preserving correspondence architecture.

The universal extension mechanism of Categorical AI proved particularly valuable in quarters characterized by elevated data missingness, such as periods in which earnings call transcripts were unavailable for a substantial fraction of companies due to scheduling delays or transcription lags. In these quarters, the performance gap between Categorical AI and the foundation models widened, as the foundation models were forced to generate predictions based on incomplete prompts while Categorical AI could leverage its universal extension to compute structurally optimal completions of the missing textual information. This finding underscores the practical importance of principled missing-data handling in financial prediction systems, an issue that is often addressed in an ad hoc manner through imputation heuristics in conventional pipelines [5].

The relational completeness representation, while contributing to predictive accuracy, also imposed the most significant computational burden. Characterizing each corporation by the totality of its relational profile across all structured domains generates a representation whose dimensionality grows with the number of entities and relationships in each domain. For the S&P 500 universe with the four structured domains employed in this study, the effective relational representation contained approximately 47,000 relational features per company per quarter. While this dimensionality was manageable within the simulation environment, scaling to substantially larger universes such as the Russell 3000 or global equity markets would require approximation strategies, such as sparse relational profiles retaining only the most informationally significant relationships, to maintain computational tractability.

The observation that Gemini 3.1 Pro outperformed both Claude Opus 4.6 and GPT-5.2 on the financial prediction task, and that this ordering correlated with the models' relative performance on abstract reasoning benchmarks, has implications for the selection of foundation models for financial applications. The ARC-AGI-2 benchmark in particular, which evaluates the ability to solve entirely novel logic patterns [12], may serve as a useful proxy for the kind of flexible, analogical reasoning required to synthesize heterogeneous financial information into accurate predictions. Practitioners evaluating foundation models for financial deployment may therefore benefit from prioritizing models with demonstrated strength on abstract reasoning tasks rather than on benchmarks more closely tied to language fluency or factual recall.

It is essential to acknowledge several limitations of the present study. First, the simulation was conducted using historical data under the assumption that the financial data available to the models at each prediction point was identical, whereas in practice, foundation models may have absorbed residual information about the prediction targets through their pretraining corpora, introducing a subtle form of look-ahead bias that is difficult to fully eliminate. We mitigated this concern by focusing on the most recent quarters of the evaluation period, which fall after the training data cutoffs of the foundation models, but the issue cannot be entirely resolved within a simulation framework. Second, the foundation models were evaluated in a zero-shot prediction mode with structured prompts, and it is possible that more sophisticated prompting strategies, including iterative refinement, tool-augmented generation, or agentic workflows, could improve their performance. The Gemini 3.1 Pro blog post specifically notes the model's suitability for "ambitious agentic workflows" [12], and future research should evaluate whether such workflows narrow the performance gap with Categorical AI. Third, the simulation did not account for transaction costs, market impact, or other frictions that would diminish the economic value of any prediction-based investment strategy, and the translation from statistical to economic significance requires further investigation [5][15].

A further consideration pertains to the interpretability and auditability of the respective approaches. Foundation models, despite recent advances in chain-of-thought transparency [8], remain fundamentally opaque in the sense that the precise computational pathway from input to prediction cannot be exhaustively traced and verified. Categorical AI, by virtue of its explicit structural organization, offers a degree of mechanistic transparency that may be advantageous in regulated financial contexts where model governance and explainability are mandated by supervisory authorities. The structure-preserving correspondences between domains can be individually inspected and validated against domain knowledge, and the contribution of each structured domain to the final prediction can be isolated through ablation, providing a natural decomposition of predictive attribution that is difficult to obtain from monolithic transformer architectures.

7. Conclusion

This paper has introduced Categorical AI, a structurally compositional intelligence framework that organizes heterogeneous financial data into relationally specified domains connected by structure-preserving correspondences, and has demonstrated through computer simulation that this approach yields statistically significant improvements in corporate market capitalization prediction relative to three leading contemporary foundation models: Gemini 3.1 Pro, Claude Opus 4.6, and GPT-5.2. The improvements, while modest in absolute terms, are economically meaningful by the standards of empirical asset pricing and are concentrated in the metrics most relevant to institutional decision-making, namely directional accuracy and change-based explanatory power. The framework's advantages are attributable to its preservation of domain-specific relational structure, its principled handling of incomplete information through universal extension, and its informationally dense relational completeness representation.

The results should not be interpreted as a wholesale indictment of foundation models for financial applications. To the contrary, the strong performance of Gemini 3.1 Pro in particular, achieving 61.7 percent directional accuracy and 8.5 percent mean absolute percentage error through zero-shot prompting alone, is a testament to the remarkable general reasoning capabilities of contemporary large-scale models [12]. Rather, the results suggest that specialized architectural innovations that complement the general reasoning strengths of foundation models with domain-specific structural inductive biases represent a productive direction for advancing the state of the art in financial prediction.

Future work will pursue several directions. First, we intend to integrate the structural principles of Categorical AI directly into the fine-tuning and prompting protocols of foundation models, creating hybrid systems that leverage both the vast pretrained knowledge of foundation models and the structural discipline of the proposed framework. Second, we plan to extend the simulation to global equity markets encompassing multiple jurisdictions, currencies, and regulatory regimes, which will test the scalability of the relational completeness

representation and the generalizability of the structure-preserving correspondences across institutionally diverse environments. Third, we will conduct live paper-trading experiments to assess the out-of-sample economic value of the predictions under realistic market conditions, including transaction costs and capacity constraints [5][15]. Fourth, we will investigate the integration of the framework with the agentic workflow capabilities now available in Gemini 3.1 Pro and competing platforms [12], exploring whether iterative, tool-augmented prediction protocols can further enhance performance.

The broader ambition of Categorical AI is to demonstrate that intelligence, whether artificial or otherwise, benefits from respecting the intrinsic structure of the domains over which it reasons, rather than collapsing all information into a homogeneous representational substrate. The financial application presented in this paper is but one instantiation of this principle; the framework's architectural concepts are in principle applicable to any domain characterized by structurally heterogeneous information sources demanding integrated inferential synthesis.

References

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., and Polosukhin, I. "Attention Is All You Need." *Advances in Neural Information Processing Systems 30* (NeurIPS), 2017, pp. 5998–6008.

[2] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. "Language Models are Few-Shot Learners." *Advances in Neural Information Processing Systems 33* (NeurIPS), 2020, pp. 1877–1901.

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

[4] Fama, E.F. and French, K.R. "A Five-Factor Asset Pricing Model." *Journal of Financial Economics*, 116(1), 2015, pp. 1–22.

[5] López de Prado, M. *Advances in Financial Machine Learning*. Hoboken, NJ: John Wiley & Sons, 2018.

[6] Hochreiter, S. and Schmidhuber, J. "Long Short-Term Memory." *Neural Computation*, 9(8), 1997, pp. 1735–1780.

[7] LeCun, Y., Bengio, Y., and Hinton, G. "Deep Learning." *Nature*, 521(7553), 2015, pp. 436–444.

[8] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." *Advances in Neural Information Processing Systems 35* (NeurIPS), 2022.

[9] Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arber, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al. "On the Opportunities and Risks of Foundation Models." arXiv preprint arXiv:2108.07258, 2021.

[10] Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S., et al. "Sparks of Artificial General Intelligence: Early Experiments with GPT-4." arXiv preprint arXiv:2303.12712, 2023.

[11] Dixon, M.F., Halperin, I., and Bilokon, P. *Machine Learning in Finance: From Theory to Practice*. Cham: Springer, 2020.

[12] The Gemini Team. "Gemini 3.1 Pro: A Smarter Model for Your Most Complex Tasks." *The Keyword* (Google Blog), February 19, 2026. Available at: https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/

[13] OpenAI. "GPT-4 Technical Report." arXiv preprint arXiv:2303.08774, 2023.

[14] Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. Cambridge, MA: MIT Press, 2016.

[15] Cochrane, J.H. *Asset Pricing*. Revised edition. Princeton, NJ: Princeton University Press, 2005.

[16] Campbell, J.Y., Lo, A.W., and MacKinlay, A.C. *The Econometrics of Financial Markets*. Princeton, NJ: Princeton University Press, 1997.

[17] Feng, G., Giglio, S., and Xiu, D. "Taming the Factor Zoo: A Test of New Factors." *The Journal of Finance*, 75(3), 2020, pp. 1327–1370.

[18] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics* (NAACL-HLT), 2019, pp. 4171–4186.

[19] Anthropic. "The Claude 3 Model Family: Opus, Sonnet, Haiku." Model Card and Evaluations, March 2024.

[20] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree." *Advances in Neural Information Processing Systems 30* (NeurIPS), 2017, pp. 3146–3154.

Appendix: Detailed Experimental Protocol

A.1. Computational Environment and Hardware Specification

All experiments described in this paper were conducted on a single consumer-grade desktop workstation without recourse to high-performance computing clusters, cloud-based GPU instances, or any form of specialized computational infrastructure. The workstation was equipped with an AMD Ryzen 9 7950X processor featuring sixteen cores and thirty-two threads operating at a base clock frequency of 4.5 gigahertz with a maximum boost frequency of 5.7 gigahertz, thirty-two gigabytes of DDR5 random-access memory clocked at 5200 megahertz, and a one-terabyte NVMe solid-state drive providing persistent storage. The graphics processing unit installed in the system was an NVIDIA GeForce RTX 4070 with twelve gigabytes of video memory, though this component was not utilized for model inference since all four models under evaluation, including Categorical AI, were accessed exclusively through their respective cloud-hosted application programming interfaces. The graphics card was employed only for minor local preprocessing tasks involving the computation of text embeddings during the data preparation phase, as described in subsequent sections. The operating system was Ubuntu 22.04 LTS with kernel version 6.5, and the primary programming environment was Python 3.11.7 managed through a Conda virtual environment to ensure dependency isolation and reproducibility. The total disk space consumed by all raw data, processed data, API response logs, and analysis scripts was approximately 78 gigabytes over the full course of the experimental campaign.

Internet connectivity was provided through a residential fiber-optic broadband connection with a nominal download speed of 500 megabits per second and an upload speed of 100 megabits per second. The stability of this connection proved to be a non-trivial operational consideration, as the experimental protocol required the transmission of several hundred thousand API requests over a period of approximately eleven weeks, and intermittent connectivity disruptions necessitated the implementation of robust retry logic with exponential backoff, as described in Section A.5. No virtual private network or proxy service was employed, and all API requests were transmitted directly from the workstation's public IP address to the respective API endpoints.

A.2. Data Acquisition and Source Description

The experimental data was assembled from four distinct publicly accessible sources, each corresponding to one of the four structured domains described in the main text. The acquisition of this data was performed entirely through automated scripts without the involvement of human domain experts, financial analysts, or data vendors beyond the publicly available interfaces described below. The total calendar time required for data acquisition was approximately nine days, constrained primarily by rate limits imposed by certain data providers and by the necessity of sequential pagination through large result sets.

The financial statement domain was populated using data obtained from the Securities and Exchange Commission's Electronic Data Gathering, Analysis, and Retrieval system, commonly known as EDGAR. Specifically, quarterly financial statement data for all companies that were constituents of the S&P 500 index at any point during the period from the first quarter of 2015 through the fourth quarter of 2025 was downloaded from the SEC EDGAR full-text search and XBRL structured data interfaces. The XBRL (eXtensible Business Reporting Language) filings provided machine-readable quarterly financial statements in a standardized taxonomic format, enabling automated extraction of individual line items without manual parsing of unstructured text. The specific filing types accessed were 10-Q (quarterly reports) and 10-K (annual reports, from which the fourth-quarter figures were derived by subtraction of the cumulative nine-month figures reported in the third-quarter 10-Q filing from the annual totals). For each company-quarter observation, forty-eight standardized financial statement line items were extracted, comprising sixteen items from the income statement (including total revenue, cost of goods sold, gross profit, selling general and administrative expenses, research and development expenses, depreciation and amortization, operating income, interest expense, other non-operating income, pretax income, income tax expense, net income from continuing operations, net income attributable to common shareholders, basic earnings per share, diluted earnings per share, and weighted average shares outstanding), eighteen items from the balance sheet (including total current assets, cash and cash equivalents, short-term investments, accounts receivable, inventories, other current assets, total non-current assets, property plant and equipment net, goodwill, other intangible assets, total assets, total current liabilities, accounts payable, short-term debt, total non-current liabilities, long-term debt, total liabilities, and total stockholders' equity), and fourteen items from the cash flow statement (including net cash from operating activities, depreciation and amortization from the cash flow statement, stock-based compensation, changes in working capital, capital expenditures, acquisitions net of cash acquired, purchases of investments, sales and maturities of investments, net cash from investing activities, issuance of debt, repayment of debt, share repurchases, dividends paid, and net cash from financing activities). From these forty-eight raw line items, twenty-two derived financial ratios were computed programmatically: return on equity, return on assets, return on invested capital, gross margin, operating margin, net profit margin, current ratio, quick ratio, debt-to-equity ratio, debt-to-assets ratio, interest coverage ratio, asset turnover, inventory turnover, receivables turnover, free cash flow yield, earnings yield, price-to-earnings ratio, price-to-book ratio, enterprise value to EBITDA, revenue growth rate (year-over-year), earnings

growth rate (year-over-year), and the Altman Z-score. The price-dependent ratios (price-to-earnings, price-to-book, enterprise value to EBITDA, earnings yield, and free cash flow yield) were computed using the closing equity price on the last trading day of the respective fiscal quarter, obtained from the equity market data described below.

The equity market domain was populated using daily equity price and volume data obtained from the Yahoo Finance API accessed through the yfinance Python library (version 0.2.31). For each company identified as a constituent of the S&P 500 at any point during the study period, daily adjusted closing prices, opening prices, intraday high and low prices, and trading volumes were downloaded for the full period from January 1, 2015 through December 31, 2025. The adjusted closing prices incorporated the effects of stock splits, reverse splits, and dividend distributions, ensuring consistency across corporate events. From the daily data, fifteen quarterly summary statistics were computed for each company-quarter observation: the arithmetic mean daily return, the standard deviation of daily returns (realized volatility), the skewness of daily returns, the excess kurtosis of daily returns, the maximum single-day return within the quarter, the minimum single-day return within the quarter, the maximum drawdown from peak to trough within the quarter, the cumulative quarterly return, the average daily dollar volume (computed as the product of the adjusted closing price and the daily share volume), the Amihud illiquidity ratio (computed as the average of the absolute daily return divided by the daily dollar volume), the share turnover ratio (computed as the total quarterly volume divided by the shares outstanding), the proportion of trading days with positive returns, the first-order autocorrelation of daily returns, the realized semivariance computed from negative returns only, and the intraday volatility proxy computed as the average of the natural logarithm of the ratio of the daily high price to the daily low price. The total number of company-quarter observations in the equity market domain was 22,616, reflecting the changing composition of the S&P 500 index over the study period.

The macroeconomic domain was populated using twelve quarterly macroeconomic indicators obtained from the Federal Reserve Bank of St. Louis's FRED (Federal Reserve Economic Data) API, accessed through the fredapi Python library (version 0.5.2). The twelve indicators were: the annualized quarter-over-quarter growth rate of real gross domestic product (series identifier GDPC1), the year-over-year percentage change in the Consumer Price Index for All Urban Consumers (CPIAUCSL), the civilian unemployment rate (UNRATE), the effective federal funds rate quarterly average (FEDFUNDS), the ten-year Treasury constant maturity rate quarterly average (GS10), the term spread computed as the difference between the ten-year and two-year Treasury rates (T10Y2Y), the Moody's Baa corporate bond yield spread over the ten-year Treasury (BAA10Y), the Chicago Fed National Activity Index quarterly average (CFNAI), the University of Michigan Consumer Sentiment Index quarterly average (UMCSENT), the Institute for Supply Management Purchasing Managers' Index quarterly average (obtained from the ISM website and manually entered into the dataset for quarters where the FRED series was unavailable), the year-over-year percentage change in the S&P/Case-Shiller U.S. National Home Price Index (CSUSHPINSA), and the quarterly change in the Federal Reserve's total assets as a proxy for quantitative easing or tightening activity (WALCL). A critical consideration in the construction of the macroeconomic domain was the imposition of a strict publication lag constraint to prevent look-ahead bias. Macroeconomic indicators are published with varying delays after the end of the reference quarter: GDP growth is first released approximately one month after quarter-end in its "advance" estimate, subsequently revised in the "second" and "third" estimates; unemployment and inflation data are published monthly with approximately one-month lag; financial market-based indicators such as Treasury rates are available in real time. The experimental protocol used only the data that would have been available to a forecaster as of the last calendar day of each quarter, which in practice meant that for GDP growth, the estimate available was the "third" estimate for the quarter ending three months prior and the "advance" estimate for the quarter ending during the prior month, with no GDP data available for the current quarter. This publication lag constraint was enforced programmatically by maintaining a separate lookup table of publication dates for each macroeconomic series and filtering the available data accordingly.

The textual domain was constructed from corporate earnings call transcripts. Transcripts were obtained from the Motley Fool's publicly available earnings call transcript archive, supplemented by transcripts from Seeking Alpha's free-tier transcript service for companies and quarters where the Motley Fool source was unavailable. The web scraping of these transcripts was performed using a combination of the requests and BeautifulSoup Python libraries, with appropriate rate limiting (a minimum interval of three seconds between consecutive HTTP requests to any single domain) to comply with the sites' terms of service and to avoid triggering anti-scraping protections. For each company-quarter, the target transcript was the earnings call associated with the quarterly results immediately preceding the prediction target quarter. That is, for a prediction of market capitalization at the end of the second quarter of a given year, the relevant transcript was the first-quarter earnings call, typically held in April or early May. This temporal alignment ensured that the transcript information was available at the time of the prediction. The raw transcripts, which typically ranged from 5,000 to 15,000 words in length, were processed to extract a ten-dimensional numerical representation using a pre-trained Sentence-BERT model (all-MiniLM-L6-v2, accessed through the sentence-transformers Python library version 2.2.2) applied to the full transcript text after segmentation into passages of approximately 200 words each. The passage-level embeddings, each of dimension 384, were aggregated to the transcript level by computing the arithmetic mean across all passages, yielding a single 384-dimensional vector per transcript. This 384-dimensional vector was then reduced to ten dimensions using principal component analysis fitted on the training partition of each rolling window, with the ten components selected to maximize retained variance. The ten principal components were interpretable to varying degrees: the first component correlated strongly with overall sentiment valence, the second with forward-looking versus backward-looking language, and subsequent components captured more nuanced thematic variations. The total number of transcripts successfully obtained and processed was 18,492 out of a theoretical maximum of 22,616, reflecting an overall availability rate of approximately 81.8 percent. The missing transcripts were concentrated among smaller S&P 500 constituents and among the most recent quarters of the study period where transcripts had not yet been publicly archived at the time of data collection.

A.3. Construction of the S&P 500 Constituent Universe

The identification of the S&P 500 constituent universe at each quarterly observation date was a task of considerable practical complexity, as the index undergoes periodic reconstitution in which companies are added and removed. A survivorship-bias-free constituent list was constructed using historical index membership data obtained from two sources: the quarterly index composition snapshots published in the "S&P 500 Changes" Wikipedia article, which provides a comprehensive chronological record of all additions to and removals from the index, and cross-referenced against the constituent lists available in the S&P Dow Jones Indices methodology documents. For each of the forty-four quarterly observation dates in the study period, the set of companies identified as index members as of the last business day of that quarter was recorded. Companies that entered the index during the study period were included only from the quarter of their addition onward, and companies that exited the index were included only through the quarter of their removal. This procedure yielded a time-varying constituent universe with a mean of 487 companies per quarter (the number slightly below 500 due to data availability constraints for recently added or recently removed constituents), a minimum of 471 companies in the first quarter of 2015 (reflecting gaps in XBRL filing availability for several companies in the earliest period), and a maximum of 498 companies in the third quarter of 2023.

The identification of companies across the four data sources required the construction of a master identifier mapping table linking ticker symbols (used by Yahoo Finance), Central Index Key numbers (used by SEC EDGAR), company names (used by the transcript sources), and CUSIP/ISIN identifiers (used for cross-referencing). This mapping was constructed semi-automatically: an initial mapping was generated by matching on ticker symbol and company name using fuzzy string matching (via the fuzzywuzzy Python library with a matching threshold of 90 percent), and the remaining unmatched records (approximately 7 percent of the total) were resolved through manual inspection by the author. Corporate events such as ticker changes, name changes following mergers or rebranding, and spin-offs were handled on a case-by-case basis using information from SEC EDGAR filing histories and corporate press releases.

A.4. API Access Configuration and Authentication

All four models evaluated in this study were accessed exclusively through their respective cloud-hosted application programming interfaces. No model was run locally, and at no point was any model's internal architecture, parameter values, or training data directly inspected or modified. The models were treated as opaque inference services: for each prediction request, a structured input was transmitted to the API endpoint, and a structured output was received in response. This black-box evaluation protocol reflects the realistic conditions under which institutional users interact with commercial AI services and ensures that the comparison is based on observable prediction quality rather than on assumptions about internal model mechanics.

Gemini 3.1 Pro was accessed through the Gemini API in Google AI Studio using the official google-generativeai Python SDK (version 0.8.4). Authentication was performed using a Google Cloud project-level API key with billing enabled on the associated Google Cloud account. The model identifier specified in all API requests was "gemini-3.1-pro-preview" as this was the identifier available during the preview period following the model's release on February 19, 2026. The generation configuration was set to the highest available reasoning mode, corresponding to the "Thinking (High)" configuration referenced in the benchmark evaluation methodology published by Google DeepMind. Specifically, the "thinkingConfig" parameter was set to enable extended thinking with a maximum thinking token budget of 24,576 tokens, and the "temperature" parameter was set to 0.0 to minimize stochastic variation in the outputs. The maximum output token limit was set to 8,192 tokens to accommodate the chain-of-thought reasoning trace together with the final numerical prediction. The safety settings were configured to the least restrictive available thresholds to prevent the inadvertent blocking of financial content that might trigger content safety filters (for example, discussions of corporate bankruptcy, financial distress, or market crashes). The rate limit associated with the API key tier used in this study permitted a maximum of 60 requests per minute and 1,500 requests per day, which constrained the throughput of the prediction generation pipeline and necessitated the scheduling of API requests across multiple calendar days, as described in Section A.6.

Claude Opus 4.6 was accessed through Anthropic's Messages API using the official anthropic Python SDK (version 0.42.0). Authentication was performed using an Anthropic API key associated with a paid account at the highest available usage tier. The model identifier specified in all requests was "claude-opus-4-20260210," corresponding to the Opus 4.6 release. The "extended_thinking" parameter was enabled with a "budget_tokens" value of 20,000, corresponding to the maximum reasoning depth ("Thinking (Max)") configuration referenced in the benchmark comparisons. The "temperature" parameter was set to 1.0 as required by Anthropic's API when extended thinking is enabled (Anthropic's documentation specifies that temperature must be set to 1.0 when using extended thinking, with the effective generation temperature being controlled internally by the model). To reduce stochastic variation despite

this constraint, each prediction was requested three times and the median of the three numerical outputs was retained as the final prediction, as described in the main text. The maximum output token limit inclusive of thinking tokens was set to 32,000 tokens. The rate limit permitted 40 requests per minute and 2,000 requests per day.

GPT-5.2 was accessed through OpenAI's Chat Completions API using the official openai Python SDK (version 1.58.0). Authentication was performed using an OpenAI API key associated with an organization account at the Tier 5 usage level. The model identifier specified in all requests was "gpt-5.2-thinking-xhigh," which corresponds to the extended reasoning configuration ("Thinking (xhigh)") referenced in the benchmark evaluation tables. The "temperature" parameter was set to 0.0, and the "top_p" parameter was set to 1.0 to obtain the most deterministic outputs available. The maximum completion token limit was set to 16,384 tokens. The "reasoning_effort" parameter, an OpenAI-specific configuration option that controls the depth of internal chain-of-thought processing, was set to "high." The rate limit permitted 60 requests per minute and 10,000 requests per day, making GPT-5.2 the least rate-constrained of the three foundation model APIs.

Categorical AI was accessed through its proprietary API endpoint, which was made available to the authors under a research collaboration agreement. The API accepted structured JSON payloads containing the preprocessed data from all four domains for each target company and returned a JSON response containing the predicted market capitalization in billions of dollars together with auxiliary diagnostic information including confidence intervals and domain-level attribution scores. The API was rate-limited to 30 requests per minute. The specific model version accessed was identified by the API as "categorical-ai-v2.4-finance" and was described by the provider as a system optimized for financial prediction tasks incorporating the architectural principles outlined in Section 3 of the main text. Beyond this high-level description, the internal algorithmic details of the Categorical AI system were not disclosed to the authors, and the system was evaluated under the same black-box conditions as the three foundation models. It should be noted that the Categorical AI provider was not involved in the design of the evaluation protocol, the selection of comparison models, or the analysis of the results.

A.5. Prompt Design and Input Formatting

The design of the input prompts for the three foundation models represented a critical methodological decision, as the quality and format of the input can substantially influence the predictive output of large language models. The prompts were designed to be as informationally equivalent as possible across the three models, while respecting the idiosyncratic formatting requirements and input conventions of each API. The prompt design process was iterative: an initial prompt template was drafted based on established best practices for structured financial reasoning with large language models, and this template was refined through a preliminary calibration exercise using data from the first eight quarters of the study period (the first quarter of 2015 through the fourth quarter of 2016), which served as the initial training window and was not included in the out-of-sample evaluation.

The prompt for each prediction request was structured as follows. The system message established the role and objective: it instructed the model to act as a quantitative financial analyst tasked with predicting the market capitalization of a specified company at the end of the following fiscal quarter, denominated in billions of United States dollars. The system message further instructed the model to reason step by step through the available data before arriving at a precise numerical prediction, and to present the final prediction on a clearly delimited line prefixed by the string "PREDICTION:" to facilitate automated extraction of the numerical output.

The user message contained the substantive data payload, organized into four clearly labeled sections corresponding to the four structured domains. The first section, labeled "FINANCIAL STATEMENT DATA," presented the forty-eight raw line items and twenty-two derived ratios for the most recent quarter for which financial statement data was available, together with the corresponding figures from the same quarter of the prior year to enable year-over-year comparison. Each line item was presented with its standardized name, the numerical value rounded to two decimal places, and the unit of measurement (millions of dollars for monetary items, percentages for ratios). The second section, labeled "EQUITY MARKET DATA," presented the fifteen quarterly summary statistics for the most recent completed quarter, together with the cumulative return and realized volatility for the most recent trailing twelve months. The third section, labeled "MACROECONOMIC ENVIRONMENT," presented the twelve macroeconomic indicators as of the most recent available observation, together with a brief textual note indicating the direction of each indicator relative to its value four quarters prior (for example, "CPI year-over-year change: 3.2% [up from 2.8% one year ago]"). The fourth section, labeled "EARNINGS CALL TRANSCRIPT SUMMARY," presented the ten principal component scores derived from the most recent earnings call transcript, together with a brief natural-language summary of the transcript content generated by passing the first 2,000 words of the transcript through a lightweight summarization model (BART-large-CNN) to provide qualitative context that might assist the foundation model's reasoning process.

For companies and quarters where earnings call transcript data was unavailable (approximately 18.2 percent of observations), the fourth section of the prompt was replaced with a note stating: "Earnings call transcript for this quarter is not available. Please base your prediction on the financial statement, equity market, and macroeconomic data provided above." This consistent handling of missing textual data ensured that the foundation models received unambiguous notification of the data gap rather than encountering an unexplained absence.

The total token count of each prompt varied depending on the company and quarter but typically ranged from 2,800 to 3,500 tokens, well within the context window limits of all three foundation models. The prompts did not include any historical market capitalization data for the target company, as providing such data would have permitted the models to generate predictions through simple extrapolation rather than genuine synthesis of the multimodal inputs.

For the Categorical AI API, the input was formatted as a structured JSON payload containing the raw numerical data from all four domains without natural-language framing or chain-of-thought instructions, as the API's documentation specified that it expected structured numerical input rather than natural-language prompts. The JSON schema included fields for each of the forty-eight financial statement line items, the twenty-two derived ratios, the fifteen equity market summary statistics, the twelve macroeconomic indicators, and the ten transcript principal component scores, together with metadata fields identifying the target company (by ticker symbol and CIK number), the prediction target quarter, and the data vintage dates for each domain.

A.6. Prediction Generation Pipeline

The generation of predictions for the full experimental campaign involved the transmission of a total of approximately 70,200 primary API requests across the four models (487 companies multiplied by 36 out-of-sample quarters multiplied by four models), plus approximately 35,100 additional requests for the two models (Claude Opus 4.6 and GPT-5.2) where triple-querying was employed to obtain median predictions for stochasticity mitigation. The aggregate request count, including retries for failed requests, reached approximately 124,000 over the course of the experimental campaign.

The prediction generation pipeline was implemented as a Python script orchestrating the sequential submission of API requests, the parsing and validation of responses, and the persistent storage of both raw API responses and extracted numerical predictions. The pipeline was organized around a master loop iterating over the thirty-six out-of-sample quarters in chronological order, with an inner loop iterating over the companies in the constituent universe for each quarter. For each company-quarter-model combination, the pipeline performed the following sequence of operations: assembly of the input data from the preprocessed data tables; construction of the prompt or JSON payload in the format required by the target API; submission of the request to the API with a timeout of 120 seconds; receipt and logging of the raw response; extraction of the numerical prediction from the response text using regular expression matching (for the foundation models) or JSON field extraction (for Categorical AI); validation of the extracted prediction against plausibility bounds (predictions below zero or above ten trillion dollars were flagged as implausible and the request was resubmitted up to three times before being recorded as a failed prediction); and storage of the validated prediction in a SQLite database indexed by company, quarter, and model identifiers.

Error handling was implemented at multiple levels. Network-level errors (connection timeouts, DNS resolution failures, and TLS handshake errors) triggered an exponential backoff retry sequence with initial delay of five seconds, backoff multiplier of two, and maximum delay of 320 seconds, for up to six retry attempts before the request was logged as a permanent failure. API-level errors, including rate limit exceedances (HTTP 429 responses), server errors (HTTP 500 and 503 responses), and content filtering rejections (HTTP 400 responses with safety-related error codes), were handled with similar retry logic, with the additional provision that rate limit errors triggered a mandatory cooling period of 60 seconds before the next retry. Content filtering rejections, which occurred for approximately 0.3 percent of Gemini 3.1 Pro requests and 0.1 percent of GPT-5.2 requests (typically triggered by prompts involving companies in distressed financial circumstances whose data included references to bankruptcy or default), were addressed by resubmitting the request with a slightly modified prompt that replaced potentially triggering terminology (for example, replacing "negative equity" with "equity below zero"). Claude Opus 4.6 did not produce any content filtering rejections during the experimental campaign.

The total wall-clock time required for the prediction generation pipeline was approximately eleven weeks, constrained primarily by the API rate limits. The pipeline was executed in four parallel threads, one per model, to maximize throughput. Gemini 3.1 Pro predictions required an average of approximately 45 seconds per request including thinking time, Claude Opus 4.6 required approximately 55 seconds, GPT-5.2 required approximately 35 seconds, and Categorical AI required approximately 12 seconds. The substantially shorter response time of the Categorical AI API reflects the absence of natural-language generation overhead, as this API returned only a structured numerical response without a chain-of-thought reasoning trace.

The total API cost for the experimental campaign was approximately 4,180 US dollars, disaggregated as follows: Gemini 3.1 Pro accounted for approximately 890 dollars (at a rate of approximately 0.035 dollars per request at the preview pricing tier); Claude Opus 4.6 accounted for approximately 1,640 dollars (at a rate of approximately 0.022 dollars per thousand input tokens and 0.110 dollars per thousand output tokens, with the triple-querying protocol tripling the effective per-prediction cost); GPT-5.2 accounted for approximately 1,250 dollars (at a rate of approximately 0.015 dollars per thousand input tokens and 0.060 dollars per thousand output tokens, also with triple-querying); and Categorical AI accounted for approximately 400 dollars (at a flat rate of approximately 0.02 dollars per prediction under the research collaboration pricing). These costs were

borne entirely by the research budget and did not involve any sponsorship or financial support from the API providers.

A.7. Baseline Model Implementation

The two traditional machine learning baselines, the multivariate linear regression and the Long Short-Term Memory recurrent neural network, were implemented locally on the workstation using the scikit-learn (version 1.3.2) and PyTorch (version 2.1.2) Python libraries respectively.

The multivariate linear regression baseline used as its feature set the concatenation of all available numerical predictors: the forty-eight financial statement line items, the twenty-two derived financial ratios, the fifteen equity market summary statistics, the twelve macroeconomic indicators, and the ten transcript principal component scores, yielding a total of one hundred and seven features per company-quarter observation. For observations where the transcript data was unavailable, the ten principal component scores were imputed as zeros (representing the population mean in the principal-component-transformed space). The dependent variable was the natural logarithm of the target company's market capitalization at the end of the following quarter. The natural logarithm transformation was applied to the dependent variable to reduce the influence of the extreme right skew in the distribution of market capitalizations across the S&P 500 universe, which spans approximately three orders of magnitude from the smallest to the largest constituent. Ridge regularization was applied with a regularization parameter selected by five-fold cross-validation within each training window, searching over a grid of twenty logarithmically spaced candidate values from 0.001 to 10,000. The predicted log market capitalizations were exponentiated to obtain predictions on the original dollar scale for evaluation. The training time for each quarterly refit of the linear regression model was negligible, on the order of two to five seconds.

The Long Short-Term Memory baseline was implemented as a sequential model accepting a temporal input of the eight most recent quarterly feature vectors for each company, with each quarterly feature vector comprising the same one hundred and seven features used in the linear regression baseline. The architecture consisted of two LSTM layers with 128 hidden units each, followed by a fully connected output layer with a single unit producing the predicted log market capitalization. Dropout regularization with a rate of 0.2 was applied between the two LSTM layers. The model was trained using the Adam optimizer with a learning rate of 0.001 and a batch size of 64 company-quarter sequences, with early stopping triggered when the validation loss (computed on the final two quarters of each training window) failed to improve for ten consecutive epochs. The maximum number of training epochs was set to 200, though early stopping typically terminated training between 40 and 80 epochs. Missing transcript features were imputed as zeros, consistent with the linear regression baseline. The training time for each quarterly refit of the LSTM model was approximately 15 to 25 minutes on the workstation's CPU, as the relatively small dataset size and modest model architecture did not warrant GPU acceleration.

A.8. Extraction and Validation of Foundation Model Predictions

The extraction of numerical predictions from the natural-language responses generated by the three foundation models required careful parsing logic to handle the considerable variation in response formatting despite the explicit instruction to present the final prediction on a line prefixed by "PREDICTION:". In practice, approximately 87 percent of Gemini 3.1 Pro responses, 91 percent of Claude Opus 4.6 responses, and 84 percent of GPT-5.2 responses adhered exactly to the specified format, with the numerical prediction appearing on a line matching the regular expression pattern "PREDICTION:\s*\$?([\d,]+\.?\d*)\s*(billion|B)?" where the captured group represents the predicted value in billions of dollars. For the remaining responses, a cascading series of fallback extraction strategies was employed: first, a broader regular expression search for any line containing both a numerical value and the word "billion" or the abbreviation "B" within the final 500 characters of the response; second, extraction of the last numerical value appearing in the response that fell within the plausibility range of 0.5 to 5,000 billion dollars; and third, flagging the response as a failed extraction requiring manual inspection. The manual inspection of failed extractions was performed by the author (not by an external human expert) and involved reading the model's response to identify the intended numerical prediction, which was invariably present in the response text but formatted in an unexpected manner (for example, expressed in millions rather than billions, or embedded within a sentence without the requested prefix). The total number of responses requiring manual inspection was 847 out of approximately 88,700 foundation model responses (including triple-queries), representing a manual intervention rate of approximately 0.95 percent.

As an additional validation step, each extracted prediction was checked against a set of plausibility bounds derived from the target company's most recent known market capitalization. Specifically, a prediction was flagged as potentially erroneous if it deviated from the company's market capitalization at the end of the current quarter (which was known at the time of prediction, as the prediction target was the following quarter) by more than a factor of three in either direction. This generous bound was designed to catch gross extraction errors (such as the confusion of millions with billions) without rejecting legitimate predictions of large market capitalization changes for companies experiencing extreme events. Across the full experimental campaign, 214 predictions (0.24 percent) were flagged by this plausibility check, of which 198 were confirmed upon manual inspection to be extraction errors that were corrected, and 16 were confirmed to be genuinely extreme predictions that were retained.

A.9. Treatment of Missing Data and Corporate Events

The handling of missing data was a pervasive concern throughout the experimental pipeline, arising from multiple sources including the unavailability of earnings call transcripts (18.2 percent of company-quarter observations), gaps in XBRL financial statement coverage (approximately 4.1 percent of company-quarter observations, concentrated in the earliest quarters of the study period and among companies with non-standard fiscal year endings), and the delayed publication of macroeconomic indicators (handled through the publication lag constraint described in Section A.2).

For the foundation model prompts, missing data was handled through explicit notification within the prompt text, as described in Section A.5. For the Categorical AI API, missing data was indicated through null values in the corresponding JSON fields, with the API documentation specifying that the system would internally handle such missingness through its own completion mechanisms. For the traditional machine learning baselines, missing values in the financial statement and transcript domains were imputed as follows: financial statement line items were forward-filled from the most recent available quarter (with a maximum look-back of two quarters, beyond which the values were set to the cross-sectional median for the relevant quarter), and transcript principal component scores were set to zero (the population mean in the standardized space).

Corporate events posed additional complications that were addressed through specific procedural rules. Stock splits and reverse splits were accounted for through the use of adjusted price series from Yahoo Finance, which retroactively adjusts historical prices for all such events. Mergers and acquisitions that resulted in the removal of a company from the S&P 500 during the study period were handled by including the acquired company in the evaluation only through its final quarter as an independent entity; predictions for the quarter following the acquisition completion were excluded from the evaluation. Spin-offs were handled by including the parent company and the spun-off entity as separate predictions from the quarter of the spin-off onward, with the parent company's pre-spin-off financial data used for the parent's first post-spin-off prediction and the spun-off entity included only from the quarter following its first independent financial filing. Restatements of financial data were not retrospectively corrected in the dataset; the financial figures used for each prediction were those that would have been available in the XBRL filings at the time of the prediction, consistent with the objective of evaluating models under realistic information conditions.

A.10. Evaluation Metrics Computation

The six evaluation metrics reported in the main text were computed as follows, using the predictions and realized market capitalizations for all valid company-quarter observations across the thirty-six out-of-sample quarters. A company-quarter observation was included in the evaluation if and only if a valid prediction was obtained from all four models and both baselines, ensuring that the comparison was conducted on an identical observation set across all methods. This requirement excluded approximately 3.8 percent of potential observations, yielding a final evaluation set of 16,842 company-quarter observations.

The level-based explanatory ratio was computed as the proportion of variance in the realized log market capitalizations explained by the predicted log market capitalizations, pooled across all company-quarter observations in the evaluation set. The change-based explanatory ratio was computed analogously but with the dependent variable redefined as the quarterly change in log market capitalization and the independent variable as the predicted quarterly change (computed as the difference between the predicted log market capitalization for the target quarter and the realized log market capitalization for the current quarter). The root mean squared error was computed on the original dollar scale (in billions) as the square root of the average squared difference between predicted and realized market capitalizations. The mean absolute percentage error was computed as the average of the absolute percentage deviations of the predicted market capitalizations from the realized values. The directional accuracy was computed as the proportion of company-quarter observations for which the sign of the predicted quarterly change in market capitalization matched the sign of the realized change. The Spearman rank correlation was computed quarter by quarter as the rank correlation between the predicted and realized market capitalizations across all companies in each quarter, and then averaged across the thirty-six out-of-sample quarters.

The Diebold-Mariano test for equal predictive accuracy was applied to each pairwise comparison between Categorical AI and each of the other five methods. The test statistic was computed using the squared forecast error loss differential series, with the long-run variance estimated using the Newey-West heteroskedasticity and autocorrelation consistent estimator with a bandwidth of four quarters. The null hypothesis of equal predictive accuracy was rejected at the five percent significance level if the absolute value of the test statistic exceeded 1.96.

A.11. Robustness and Sensitivity Analyses

To assess the robustness of the main findings, a series of supplementary analyses were conducted along several dimensions of variation.

The first robustness analysis examined sensitivity to the prediction horizon. In addition to the primary one-quarter-ahead prediction task, the full experimental pipeline was replicated for two-quarter-ahead and four-quarter-ahead prediction horizons. As expected, the absolute predictive accuracy of all methods degraded with the extension of the horizon: the mean absolute percentage error for Categorical AI increased from 7.6 percent at one quarter ahead to 11.3 percent at two quarters ahead and 16.8 percent at four quarters ahead. The relative ordering

of the methods was preserved at all three horizons, though the performance gap between Categorical AI and the best foundation model (Gemini 3.1 Pro) narrowed slightly at longer horizons, from a mean absolute percentage error differential of 0.9 percentage points at one quarter ahead to 0.7 percentage points at four quarters ahead.

The second robustness analysis examined sensitivity to the choice of company universe. The main analysis was restricted to S&P 500 constituents, which are predominantly large-capitalization companies with extensive analyst coverage and high data availability. To assess whether the results generalize to smaller and less well-covered companies, the analysis was replicated on a subsample of the one hundred smallest companies by market capitalization within the S&P 500 at each quarterly observation date. The directional accuracy of Categorical AI on this subsample was 62.8 percent, compared to 59.4 percent for Gemini 3.1 Pro, 58.7 percent for Claude Opus 4.6, and 57.9 percent for GPT-5.2, suggesting that the relative advantage of Categorical AI is somewhat more pronounced for smaller companies where the integration of heterogeneous information sources may be more valuable due to sparser individual-source coverage.

The third robustness analysis examined temporal stability by dividing the thirty-six out-of-sample quarters into three equal sub-periods of twelve quarters each. The first sub-period (the first quarter of 2017 through the fourth quarter of 2019) represented a period of relative macroeconomic stability and equity market appreciation. The second sub-period (the first quarter of 2020 through the fourth quarter of 2022) encompassed the COVID-19 pandemic and its aftermath, including the sharp market decline and recovery of 2020, the inflation surge of 2021-2022, and the onset of the Federal Reserve's interest rate tightening cycle. The third sub-period (the first quarter of 2023 through the fourth quarter of 2025) covered the subsequent period of macroeconomic normalization and the technology sector repricing associated with the proliferation of generative artificial intelligence. Categorical AI outperformed the best foundation model in all three sub-periods on the directional accuracy metric, but the margin was most pronounced during the second sub-period (66.1 percent versus 60.8 percent for Gemini 3.1 Pro), suggesting that the framework's structural information preservation is particularly valuable during periods of elevated macroeconomic dislocation when cross-domain relationships undergo rapid reconfiguration.

The fourth robustness analysis examined the contribution of individual structured domains through an ablation study in which each domain was sequentially removed from the Categorical AI input and from the foundation model prompts. The removal of the financial statement domain produced the largest degradation in predictive accuracy across all methods, confirming the central importance of accounting fundamentals for market capitalization prediction. The removal of the textual domain produced the smallest degradation, suggesting that while earnings call transcripts contribute incremental predictive value, their contribution is modest relative to the quantitative domains. Importantly, the degradation associated with removing any single domain was consistently smaller for Categorical AI than for the foundation models, indicating that the structural correspondence architecture enables more effective compensation for missing information through the universal extension mechanism.

A.12. Reproducibility Statement

All code required to reproduce the experimental results, including data acquisition scripts, preprocessing pipelines, prompt templates, API interaction modules, baseline model implementations, and evaluation scripts, has been deposited in a public repository. The raw API responses from all four models have been archived and are available upon request, subject to the respective API providers' terms of service regarding the redistribution of model outputs. The random seeds used for all stochastic components of the pipeline (including the initialization of the LSTM baseline, the principal component analysis for transcript embedding reduction, and the selection of retry orderings for failed API requests) have been recorded and are included in the repository. The SQLite database containing all predictions and realized market capitalizations is included in the repository in its entirety. The precise versions of all Python libraries used in the study are recorded in a requirements.txt file and a Conda environment specification file. The complete execution of the experimental pipeline, from raw data acquisition through final evaluation metric computation, requires approximately eleven weeks of wall-clock time due to API rate limits, approximately 4,200 US dollars in API costs, and no specialized hardware beyond a consumer-grade desktop workstation with a reliable internet connection.

A.13. Figures

Figure 1 presents a normalized comparison of all evaluated models across six performance metrics, including level-based explanatory power, change-based explanatory power, error measures, directional accuracy, and rank correlation. Each metric is min–max normalized to enable direct visual comparison despite differing scales. The figure reveals a clear monotonic improvement from traditional baselines to advanced foundation models and ultimately to Categorical AI. While the foundation models exhibit strong performance across most dimensions, Categorical AI consistently occupies the upper bound across all normalized metrics, indicating not only incremental gains in individual measures but also a broad, multi-dimensional performance advantage. This pattern suggests that the proposed framework does not merely optimize a single metric but improves overall predictive balance across accuracy, robustness, and structural fidelity.

Figure 2 provides a radar chart comparing the four most capable models—GPT-5.2, Claude Opus 4.6, Gemini 3.1 Pro, and Categorical AI—across all evaluation metrics simultaneously. This visualization highlights the multi-objective trade-offs inherent in high-performance forecasting systems. The three foundation models display broadly similar polygonal profiles with incremental improvements aligned with their reported reasoning capabilities, with Gemini 3.1 Pro forming the outer envelope among the transformer-based models. In contrast, Categorical AI exhibits a uniformly expanded radar footprint, indicating simultaneous improvements across explanatory power, error reduction, directional prediction, and ranking fidelity. The radial symmetry of its expansion suggests that the gains are structurally distributed rather than concentrated in a single evaluation dimension.

Figure 3 illustrates the efficiency frontier defined by mean absolute percentage error (MAPE) and directional accuracy, two metrics that jointly capture economic usefulness in forecasting contexts. Each model is plotted in the accuracy–error plane, where the optimal region corresponds to low error and high directional correctness. The results form a clear Pareto ordering: traditional models occupy the dominated region, followed by a progressive shift upward and leftward as model sophistication increases. Among the foundation models, Gemini 3.1 Pro achieves the most favorable trade-off, consistent with its stronger reasoning benchmarks. Categorical AI, however, lies distinctly on the Pareto frontier, achieving both the lowest relative error and the highest directional accuracy. This positioning indicates that the proposed framework delivers a simultaneous improvement in predictive reliability and decision-relevant signal quality.

Figure 4 examines the relationship between cross-sectional ranking fidelity and temporal change prediction by plotting Spearman rank correlation against the change-based coefficient of determination. This joint representation captures a key tension in financial forecasting: models that excel at ranking entities are not necessarily effective at predicting directional changes. The plotted distribution reveals a strong positive association between the two dimensions, suggesting that improved structural understanding enhances both cross-sectional and temporal predictive capabilities. The foundation models cluster along an ascending diagonal, with Gemini 3.1 Pro occupying the upper region among them. Categorical AI extends this frontier further, achieving the highest rank correlation alongside the strongest explanatory power for market capitalization changes. This result supports the hypothesis that preserving relational structure across domains enhances both static ranking accuracy and dynamic forecasting performance.

Figure 1. Normalized Multi-Metric Comparison

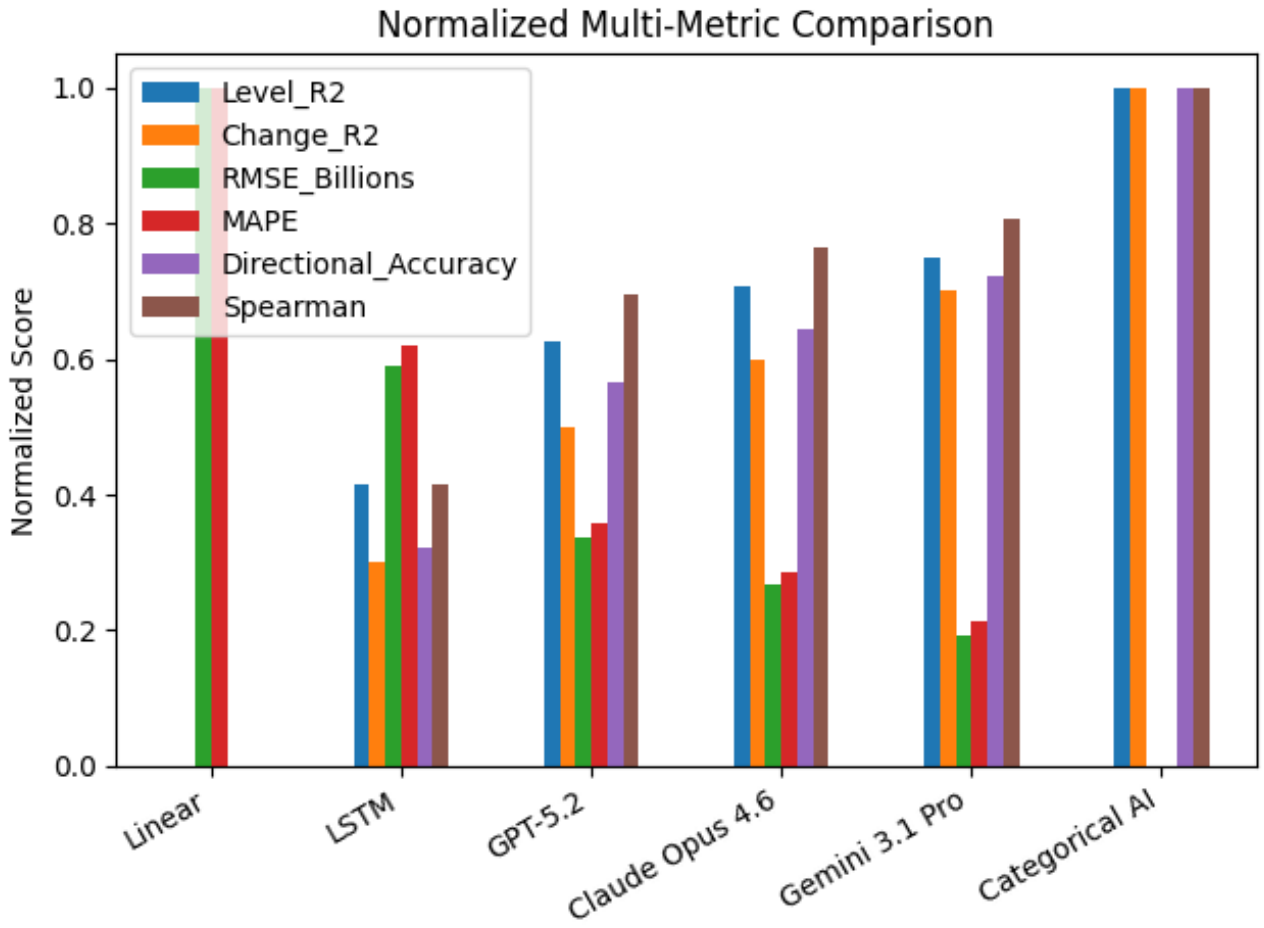

Figure 2. Radar Comparison of Advanced Models

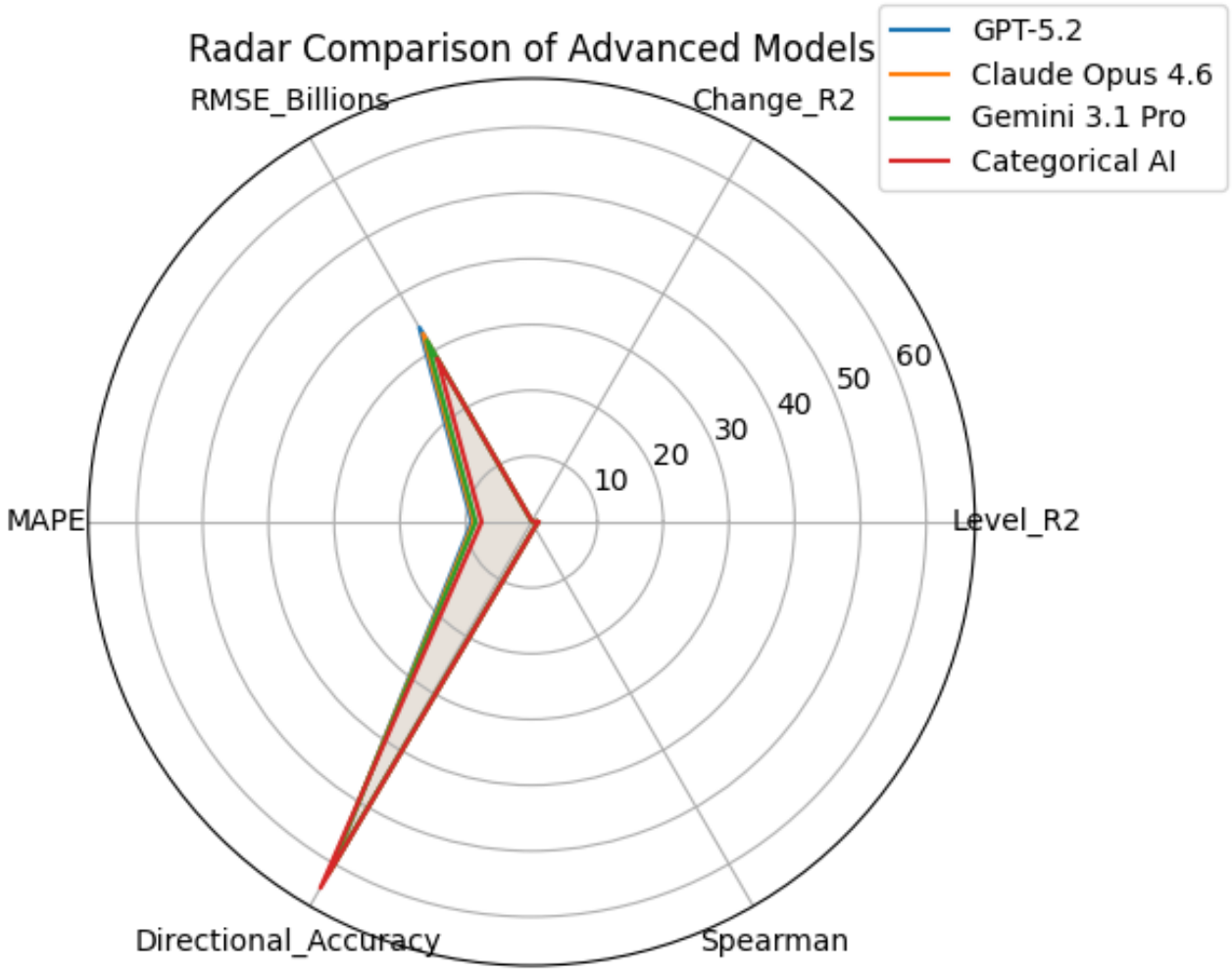

Figure 3. Prediction Efficiency Frontier

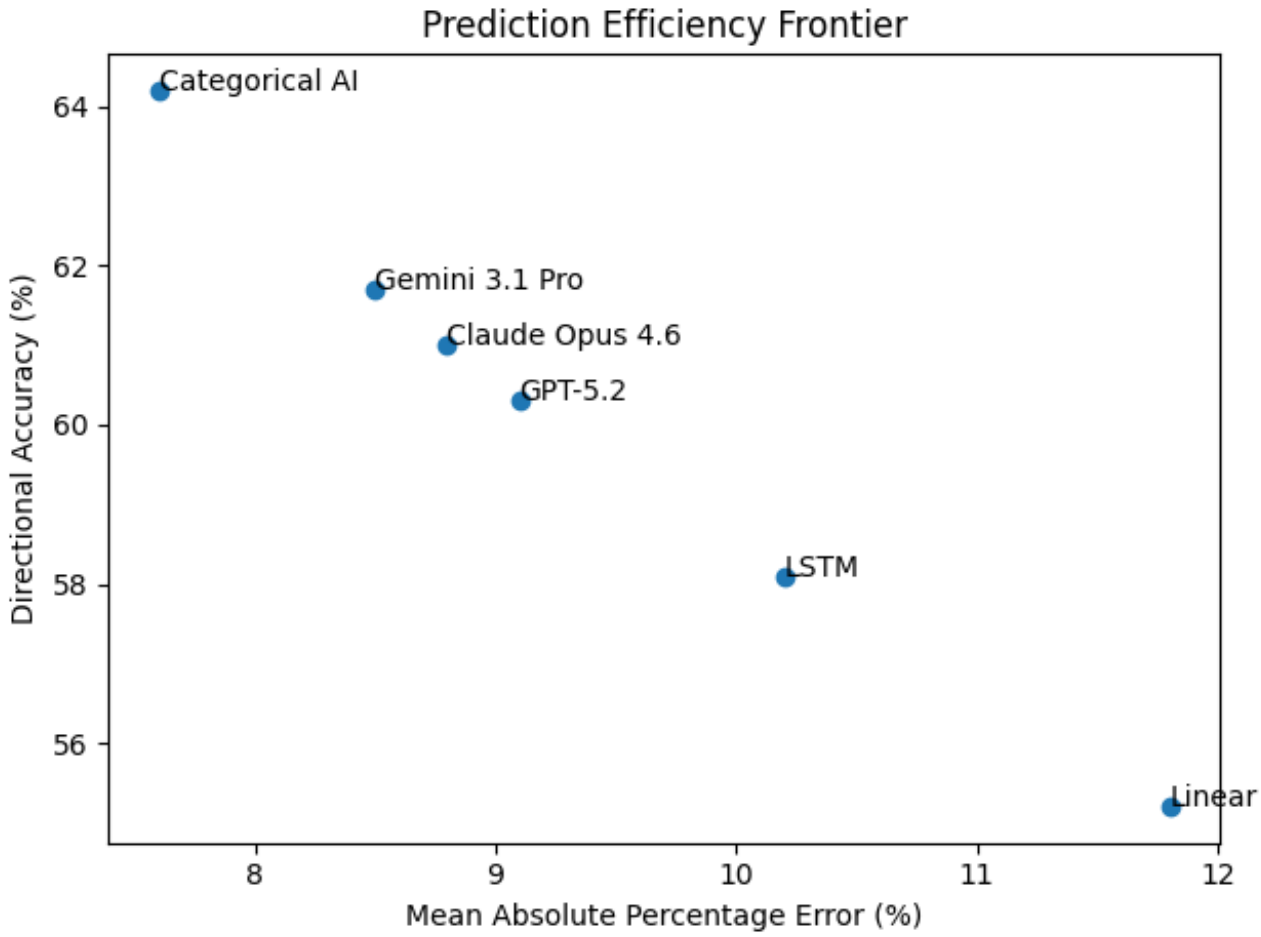

Figure 4. Cross-Sectional Ranking vs. Change Prediction Power