Causal-Enhanced and Regime-Adaptive Hybrid Framework for Asset Return Prediction

Xinze Fan, Zhongliang Yang*, Linna Zhou
School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing, China
yangzl@bupt.edu.cn

***Abstract*—Forecasting asset returns presents a formidable challenge for neural architectures, characterized by low signal-to-noise ratios, non-stationary concept drift, and the structural misalignment between semantic and numeric modalities. Existing paradigms often necessitate a compromise between interpretability and capacity, struggling to reconcile statistical rigor with the expressiveness of deep learning. To resolve these limitations, we propose a Causal-Enhanced, Regime-Adaptive Neuro-Symbolic Framework. Departing from monolithic black-box approaches, our architecture imposes a disciplined causal information bottleneck: a consensus-based structural discovery mechanism first filters spurious correlations to control the False Discovery Rate (FDR). Subsequently, a continuous regime-embedding network dynamically modulates factor attention, enabling smooth adaptation to evolving market distributions. Finally, a calibrated LLM-numeric fusion head aligns the semantic reasoning of Large Language Models with rigorous regression targets, rectifying modality-specific miscalibration. Extensive out-of-sample evaluations on U.S. equities (1926–2024) and Chinese credit bonds demonstrate that this architecture achieves superior generalization, attaining an annualized Sharpe Ratio of 1.63 and boosting out-of-sample R^2 by over 25% relative to state-of-the-art tabular foundation models. The framework exhibits robust out-of-distribution (OOD) performance, effectively mitigating downside risk during structural market turbulence.**

***Index Terms*—Causal Discovery, Regime Adaptation, Large Language Models, Asset Pricing, Modality Alignment**

I. Introduction

The central challenge in empirical asset pricing lies in distinguishing robust predictive signals from the cacophony of high-dimensional market noise. Accurate forecasting of asset returns is fundamental to capital allocation and risk management, yet it remains an elusive goal due to the adversarial nature of financial markets. Unlike physical systems governed by immutable laws, financial markets are characterized by a low signal-to-noise ratio and evolving underlying structures. Consequently, researchers face a trilemma arising from three intertwined complexities: the "factor zoo" problem, where high-dimensional data increases the risk of spurious correlations [1]–[3]; pervasive nonstationarity, where structural regime shifts and distribution drifts render static models obsolete [4], [5]; and the modality gap, where qualitative drivers embedded in textual data are often divorced from quantitative valuation models [6], [7].

Current methodologies typically address these challenges in isolation, resulting in suboptimal architectural trade-offs. Traditional econometric approaches prioritize interpretability and stability but often fail to capture the complex, non-linear dynamics inherent in modern markets. Conversely, modern deep learning paradigms, while excelling at pattern recognition, are prone to overfitting noise, particularly when market regimes shift abruptly. More recently, the emergence of Large Language Models (LLMs) has offered a promising avenue for incorporating semantic information from news and reports. However, applying LLMs directly to numeric forecasting reveals a critical limitation: these models exhibit a fundamental misalignment between their probabilistic token-generation objectives and the rigorous demands of continuous regression, often leading to miscalibration and "arithmetic hallucination" [8]–[10].

In this paper, we argue that a robust forecasting system requires an architecture structurally designed to mitigate these respective weaknesses rather than merely aggregating modalities. We introduce a *Causal-Enhanced, Regime-Adaptive Hybrid Framework* that imposes a disciplined inductive bias on the prediction process. Departing from monolithic black-box architectures that ingest raw, noisy inputs, our approach first employs a rigorous causal discovery pipeline to act as an information bottleneck, filtering spurious associations. To address nonstationarity, we discard static weighting schemes in favor of continuous regime embeddings, enabling the model to dynamically modulate factor attention based on the inferred market state. Finally, to bridge the gap between semantic reasoning and numeric precision, we design a calibration mechanism that explicitly aligns the LLM's latent representations with rigorous regression targets.

This framework explicitly targets transferability across heterogeneous market conditions while preserving economic interpretability. By integrating causal screening with regime-aware adaptation, we ensure that the model relies on economically meaningful drivers rather than transient statistical artifacts. Our contributions are summarized as follows:

- **Construction of a Causal-Driven Compact Predictor.** We move beyond correlation-based feature selection by developing a reproducible pipeline that synthesizes causal structure discovery with econometric validation. This process yields a refined factor set $\mathcal{F}_{\text{causal}}$ that strictly controls the False Discovery Rate (FDR), ensuring that subsequent neural modeling steps are grounded in robust economic relationships rather than data mining bias.
- **Dynamic Adaptation via Continuous Regime Embeddings.** Recognizing that discrete regime-switching models often lack the granularity to capture gradual market drifts, we introduce a mechanism based on con-

tinuous regime embeddings $\mathcal{R}_t = \psi(\mathbf{p}_t)$. A learnable mapping function generates time-varying factor weights w_t, enforcing temporal smoothness and consistency with the causal graph to maintain predictive utility during structural transitions.
- **Calibrated Fusion of Semantic and Numeric Intelligence.** We propose a hybrid architecture that synergizes the semantic reasoning capabilities of LLMs with the precision of numerical neural networks. By serializing weighted tabular inputs into soft prompts and attaching a specialized calibration head, we rectify the scale insensitivity and bias inherent in language models. Extensive empirical evaluations on U.S. equity data and the Chinese credit bond market demonstrate that this architecture achieves superior out-of-sample accuracy and economic performance compared to state-of-the-art baselines.

II. Related Work

The methodological landscape of asset return prediction has evolved along three distinct but converging trajectories: causal structure learning, adaptive neural architectures, and multimodal foundation models.

A. Causal Structure Learning and Econometric Inference

Early literature focused heavily on linear factor models and stringent statistical testing to identify robust predictors. Approaches such as sparse regression and structural selection emphasize parsimony and interpretability [11], [12]. While these methods provide theoretical guarantees against overfitting, they often struggle with the high dimensionality of modern datasets and the non-linear interactions characteristic of financial markets. Recent advances in causal discovery have attempted to mitigate the "factor zoo" problem by distinguishing genuine drivers from spurious correlates via directed acyclic graphs (DAGs) [13], [14]. However, standard causal algorithms (e.g., PC, GES) generally rely on the assumption of stationarity. This premise is frequently violated in financial time series, where underlying causal mechanisms drift over time, necessitating algorithms capable of distinguishing stable causal links from transient statistical correlations.

B. Neural Architectures under Distribution Shifts

To capture complex non-linearities, the field has increasingly adopted sophisticated machine learning techniques, ranging from gradient-boosted decision trees to attention-based architectures like TabNet and Tab-Transformer [15], [16]. Although these models demonstrate superior cross-sectional predictive power, they are notoriously brittle in the face of out-of-distribution (OOD) generalization. The literature on domain adaptation and regime switching [5], [17] highlights the necessity of dynamic modeling. Nevertheless, most existing solutions rely on discrete state-space models or sliding window approaches, which result in abrupt parameter jumps or lagged adaptation due to the discretization of continuous market dynamics. There remains a paucity of methods that employ continuous representations of market states to modulate factor importance smoothly, ensuring temporal consistency in neural predictions.

C. LLMs and Semantic-Numeric Alignment

The integration of unstructured textual data—news, earnings calls, and analyst reports—has long been a goal of empirical finance. Prior to the era of Large Language Models (LLMs), this was primarily achieved through lexicon-based sentiment analysis or static embeddings [6], [7]. The advent of LLMs has revolutionized this domain, enabling the extraction of deep semantic features and complex reasoning chains [18], [19]. Despite their prowess in natural language understanding, LLMs exhibit significant limitations in direct numeric forecasting tasks. Studies indicate that LLMs frequently suffer from numeric miscalibration and "arithmetic hallucination," requiring specific architectural modifications to function as reliable regressors [20]. Current hybrid approaches often resort to naive concatenation of text embeddings and tabular features. This strategy fails to explicitly align the high-dimensional semantic space of the LLM with the rigorous causal and numeric constraints required for financial decision-making. Our work addresses this specific gap by enforcing causal sparsity and numerical calibration within an LLM-centric workflow.

III. Methodology

We propose a *Causal-Enhanced, Regime-Adaptive Hybrid Framework* designed to address the structural fragility of conventional asset pricing models. As illustrated in Fig. 1, the architecture departs from monolithic black-box predictors by orchestrating a disciplined three-stage pipeline. First, to mitigate the "factor zoo" dilemma, we implement a **Causal Discovery Module** that acts as an information bottleneck, filtering out spurious correlates before they enter the prediction horizon. Second, recognizing that causal mechanisms evolve, a **Regime-Adaptive Mechanism** projects market states into continuous embeddings, dynamically modulating the influence of validated factors. Finally, to bridge the modality gap, an **LLM-Numeric Fusion Head** synthesizes quantitative signals with semantic reasoning, calibrated strictly to financial targets. This design ensures that the model remains robust under distributional shifts while preserving the economic interpretability of its decision drivers.

A. Problem Formulation under Causal Constraints

Let $r_{i,t+1}$ denote the excess return of firm i at month $t+1$. At time t, the observable predictor space is defined by the high-dimensional vector:

$$z_{i,t} = \left[m_{i,t,1}, \ldots, m_{i,t,P},\ g_{t,1}, \ldots, g_{t,D}\right]^\top, \tag{1}$$

where $m_{i,t,\cdot}$ represents firm-specific idiosyncratic features (e.g., momentum, value signals) and $g_{t,\cdot}$ denotes macroeconomic state variables. The canonical forecasting objective seeks to estimate the conditional expectation:

$$r_{i,t+1} = \mathbb{E}\left[r_{i,t+1} \mid z_{i,t}\right] + \varepsilon_{i,t+1}. \tag{2}$$

However, directly estimating Eq. (2) via standard supervised learning is prone to failure in financial contexts due to the

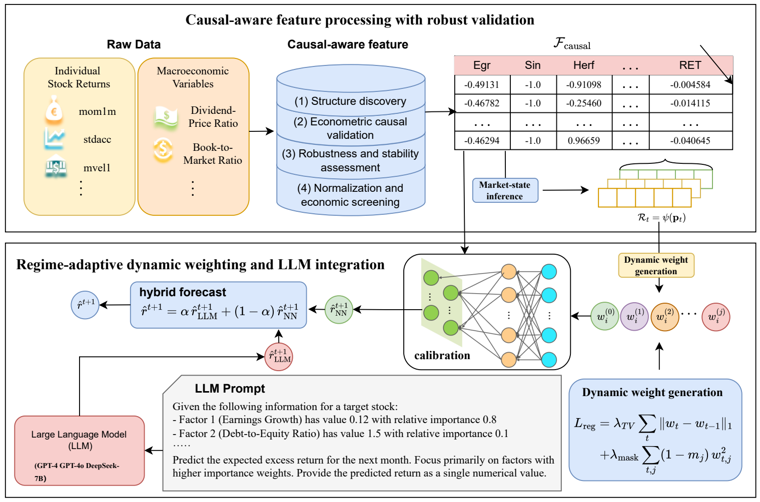

Fig. 1. Overall Architecture of the Causal-Enhanced and Regime-Adaptive LLM Framework.

presence of spurious correlations within $\mathcal{Z}$ and the non-stationarity of the mapping function. To resolve this, we impose a **Causal Information Bottleneck**. We assume that the invariant predictive signal is contained solely within a causally-validated subset $\mathcal{F}_{\text{causal}} \subset \mathcal{Z}$. Consequently, the predictive mapping is reformulated as:

$$\mathbb{E}\big[r_{i,t+1} \mid z_{i,t}\big] \approx g^*\big(\mathcal{F}_{\text{causal}}(z_{i,t}),\ \mathcal{R}_t,\ w_t\big), \tag{3}$$

where $\mathcal{F}_{\text{causal}}(\cdot)$ represents the screening operator derived from structural discovery, $\mathcal{R}_t$ is a compact, continuous embedding of the latent market state at t (detailed in Section III-C0a), and w_t is a learnable, time-varying weight vector. This formulation explicitly acknowledges that while causal factors are robust, their relative pricing power (w_t) is state-dependent, necessitating a model that is both structurally sparse and temporally adaptive.

B. Causal-Aware Feature Processing with Robust Validation

Constructing a reliable predictor set from a high-dimensional "factor zoo" requires moving beyond simple correlation metrics, which are often inflated by data mining. We propose a rigorous, reproducible pipeline designed to extract $\mathcal{F}_{\text{causal}}$—a factor set that is statistically significant, structurally stable, and economically interpretable. This process operates as a progressive filtration system through four logically distinct stages.

a) Stage 1: Consensus-Based Structure Discovery.: Relying on a single causal discovery algorithm often introduces specific inductive biases. To mitigate this, we employ a *multi-view consensus strategy*. We first utilize the continuous optimization framework of NOTEARS to enforce acyclicity on the global graph structure:

$$\min_A\ \ell(X;A) + \lambda\|A\|_1 \quad \text{s.t.} \quad \operatorname{tr}(e^{A\circ A}) - d = 0. \tag{4}$$

Recognizing that linear formulations may miss complex interactions, we extend this using MLP-NOTEARS [21]. Furthermore, to address financial time-series properties—namely non-Gaussianity and temporal autocorrelation—we complement NOTEARS with TS-LiNGAM [22] and PCMCI [23]. The final candidate set $\mathcal{F}_{\text{cand}}$ is constructed via a strict intersection rule: a variable is retained only if identified by NOTEARS and corroborated by at least one auxiliary method, significantly reducing algorithm-specific false positives.

b) Stage 2: Econometric Validation and FDR Control.: Graph-theoretical edges suggest potential causality but do not quantify predictive efficacy. To bridge this gap, we subject each candidate $f_j \in \mathcal{F}_{\text{cand}}$ to a Granger-style predictive test. We compare a restricted baseline model ($\mathcal{M}_{\text{Null}}$, where $\beta_\ell = 0$) against an unrestricted augmented model:

$$\mathcal{M}_{\text{Alt}}: \quad r_{i,t+1} = \alpha + \sum_{\ell=1}^{L}(\phi_\ell r_{i,t+1-\ell} + \gamma_\ell x_{t-\ell} + \beta_\ell f_{j,t-\ell}) + u_t. \tag{5}$$

We rigorously test the null hypothesis $H_0: \beta_\ell = 0$ using standard Wald tests. Crucially, given the high cardinality of the candidate space, we apply the Benjamini–Hochberg procedure to control the False Discovery Rate (FDR) at level q, ensuring that selected factors possess genuine incremental power rather than being statistical flukes.

c) Stage 3: Perturbation-Based Stability Assessment.: A robust factor must maintain validity across different market conditions. Factors passing the econometric filter are further scrutinized using a stability metric derived from bootstrap resampling: $\pi_j = \Pr(f_j \in \mathcal{F}_{\text{selected}} \mid \mathcal{D}_{\text{resample}})$. We discard any factor with stability $\pi_j \le \pi^\star$, thereby filtering out "fragile" signals that are sensitive to specific time windows or outliers.

d) Stage 4: Economic Sanity Screening.: Finally, statistical significance must not supersede economic logic. The surviving factors undergo qualitative screening to eliminate look-ahead bias and ensure interpretability (e.g., sign consistency with economic theory). The resulting set $\mathcal{F}_{\text{causal}}$ represents a distilled group of factors that are mathematically causal, statistically predictive, and economically actionable.

C. Regime-Adaptive Dynamic Weighting and Calibrated LLM Integration

Having established a statistically robust factor set $\mathcal{F}_{\text{causal}}$, the final challenge is to deploy these factors in a non-stationary environment where their predictive efficacy fluctuates over time. We address this through a dual-mechanism architecture: a *Regime-Aware Gating Network* that dynamically modulates factor exposure, and a *Calibrated Semantic-Numeric Fusion* module that synthesizes quantitative rigor with qualitative reasoning.

a) Continuous Market-State Embedding.: Financial markets do not switch between discrete states (e.g., "Bull" vs. "Bear") instantaneously; rather, they exhibit continuous drifts in risk premia. Standard discrete regime models (e.g., HMMs) often introduce harmful discontinuities in prediction. To capture this fluidity, we first compute the posterior state probabilities $\mathbf{p}_t$ from the macroeconomic sequence $\{g_{1:t}\}$. Crucially, we project these discrete probabilities into a continuous latent manifold via a learnable encoder ψ, yielding the regime embedding $\mathcal{R}_t = \psi(\mathbf{p}_t)$. This dense representation allows the model to interpolate between market states, enabling smooth transitions in factor allocation.

b) Causally-Enhanced Dynamic Gating.: To translate the inferred regime $\mathcal{R}_t$ into actionable decisions, we employ a parametric gating network generating a time-varying weight vector w_t:

$$w_t = \mathrm{softmax}\Big(U^\top \tanh\left(W_1 g_t + W_2 \mathcal{R}_t + b\right)\Big). \quad (6)$$

To prevent overfitting, we impose structural priors via the regularization term L_{reg}:

$$L_{\mathrm{reg}} = \lambda_{\mathrm{TV}} \sum_{t=2}^{T} \|w_t - w_{t-1}\|_1 + \lambda_{\mathrm{graph}} \sum_{t=1}^{T} \sum_{j=1}^{P} (1 - m_j) w_{t,j}^2 \quad (7)$$

The first term (Total Variation) enforces *temporal smoothness* to control turnover costs. The second enforces *causal consistency*, penalizing weights on factors rejected by the causal discovery ($m_j = 0$) to ensure the dynamic model respects the static causal graph.

c) LLM-Numeric Hybrid Fusion with Calibration.: Since standard LLMs struggle with direct regression, we treat them as semantic reasoners rather than calculators. Weighted numeric features $\mathcal{F}_{\mathrm{causal}}(z_{i,t}) \odot w_t$ are serialized via soft-token embeddings $\varphi(\cdot)$ and fed into the LLM alongside the regime context $\mathcal{R}_t$.

To correct inherent "arithmetic hallucinations" and scale insensitivity, we attach a lightweight linear **Calibration Head**. The final forecast ensembles the calibrated LLM with a specialized Neural Network (NN):

$$\hat{r}_{\mathrm{LLM}}^{t+1} = \mathrm{CalibHead}\big(\mathrm{LLM}(\varphi(\mathcal{F}_{\mathrm{causal}}(z_{i,t}) \odot w_t),\ \mathcal{R}_t)\big),$$
$$\hat{r}^{t+1} = \alpha\, \hat{r}_{\mathrm{LLM}}^{t+1} + (1-\alpha)\, \mathrm{NN}\big(\mathcal{F}_{\mathrm{causal}}(z_{i,t}) \odot w_t,\ \mathcal{R}_t\big). \quad (8)$$

The parameter α balances orthogonal signals: the NN captures non-linear numeric patterns, while the LLM provides semantic and regime context.

d) Optimization Strategy.: We train the framework to minimize $L = L_{\mathrm{pred}} + L_{\mathrm{reg}} + \lambda_{\mathrm{graph}} L_{\mathrm{graph}}$, where L_{graph} aligns attention weights with the ideal causal mask. To ensure stability, we adopt a *decoupled curriculum learning strategy*: first pretraining the numeric/gating modules to stabilize w_t, then fine-tuning the LLM prompts and calibration head. This prevents noisy LLM gradients from destabilizing the causal feature selection.

IV. Experiments

A. Implementation Details and Experimental Setup

To ensure our findings capture modern market dynamics, we extended the standard CRSP U.S. equity dataset to cover the period from January 1926 to December 2024. This expansive window includes critical recent stress scenarios, such as the COVID-19 crash and the subsequent inflationary regime shift, providing a rigorous testbed for model adaptability. The dataset comprises 94 standardized firm-level and macroeconomic predictors, with industry heterogeneity controlled via Fama–French 49 industry classification. We strictly exclude penny stocks and illiquid firms to prevent microstructure noise from inflating performance metrics.

We benchmark our framework against a hierarchy of baselines representing the evolution of asset pricing: (1) **Linear Econometric Models** (OLS, PCR) that prioritize stability but lack capacity; (2) **Non-linear Machine Learning** (Random Forests, NN3/4) that capture complexity but risk overfitting; (3) **Tabular Foundation Models** (TabNet [24], TabTransformer [25]) representing state-of-the-art deep tabular learning; and (4) **Frontier LLMs** (DeepSeek-7B, GPT-4, GPT-4o) utilized in a naive zero-shot manner.

Evaluation is conducted using a strict recursive window approach to eliminate look-ahead bias. We assess performance through two lenses: statistical accuracy (out-of-sample R^2, MSE) and economic utility (Annualized Return, Sharpe Ratio, and Maximum Drawdown), utilizing Newey–West adjusted t-statistics to verify significance.

Computational Configuration and Reproducibility. All models were implemented in PyTorch and trained on a cluster of NVIDIA A100 GPUs. Addressing concerns regarding computational complexity, we designed the heavy-computation module—Causal Structure Discovery—as a low-frequency offline process (updated annually), ensuring that the monthly inference latency remains negligible ($<$ 10ms per stock). To guarantee fair comparison, hyperparameters for all models (including baselines and ablation variants) were rigorously optimized via a rolling validation scheme (using the 12 months prior to each test window). Specifically, for the ablation study, each reduced model variant was *independently re-tuned* to its optimal configuration. As illustrated in Fig. 2, the model exhibits structural stability: the Sharpe ratio remains robust ($>$ 1.5) across a broad range of the causal sparsity penalty λ_{graph} and regime smoothness λ_{TV}, confirming that performance is driven by the architectural design rather than overfitting to narrow hyperparameter niches.

B. Experimental Results and Analysis

Table I reveals a clear stratification of performance driven by model capability. Traditional econometric methods (OLS, PCR) struggle to beat the historical mean, often yielding negligible or negative out-of-sample R^2, confirming that linear factors alone are insufficient in efficient markets. While standard machine learning models (RF, NN) unlock non-linear

TABLE I

OUT-OF-SAMPLE PREDICTION PERFORMANCE ON MONTHLY STOCK RETURNS. FOR EACH SUBSET WE REPORT R^2%, RMSE, SHARPE (ANNUALIZED) AND ANNUAL RETURN (%). BEST R^2 PER SUBSET IN **BOLD**.

Model	All				Top 1000				Bottom 1000			
	R^2%	RMSE	Sharpe	Ann. Ret.	R^2%	RMSE	Sharpe	Ann. Ret.	R^2%	RMSE	Sharpe	Ann. Ret.
Econometric												
OLS	-3.46	0.122	0.42	6.8	-11.28	0.139	0.28	5.1	-1.30	0.115	0.47	7.2
OLS-S	0.16	0.110	0.56	9.3	0.31	0.104	0.61	10.1	0.17	0.111	0.54	8.9
PCR	0.26	0.108	0.61	10.2	0.06	0.106	0.58	9.7	0.34	0.107	0.63	10.6
Machine Learning												
RF	0.33	0.105	0.66	11.2	0.63	0.101	0.72	12.1	0.35	0.106	0.65	11.0
NN3	0.40	0.102	0.72	12.3	0.70	0.099	0.77	13.0	0.45	0.104	0.70	12.1
NN4	0.39	0.101	0.74	12.6	0.67	0.098	0.79	13.2	0.47	0.103	0.72	12.4
Tabular foundation models												
TabNet	1.41	0.081	1.12	16.4	1.54	0.086	0.95	15.2	0.53	0.099	0.82	13.7
Tab-Transformer	1.24	0.087	0.91	14.8	1.97	0.062	1.62	22.7	0.64	0.092	0.87	13.8
LLM-base(Ours)												
DeepSeek-7B	1.43	0.079	1.34	18.6	1.78	0.063	1.54	21.1	1.12	0.111	1.19	17.4
GPT-4	1.62	0.078	1.51	21.1	2.08	0.061	1.71	23.6	1.23	0.081	1.36	19.9
GPT-4o	**1.72**	**0.075**	**1.63**	**22.7**	**2.76**	**0.052**	**1.83**	**25.2**	**1.33**	**0.092**	**1.48**	**21.5**

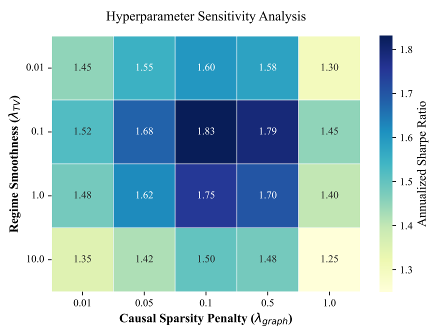

Fig. 2. **Hyperparameter Sensitivity Analysis.** The heatmap illustrates the out-of-sample Sharpe Ratio across varying causal sparsity (λ_{graph}) and regime smoothness (λ_{TV}) penalties. The broad "high-performance plateau" (Sharpe > 1.6) confirms the model's structural robustness to parameter tuning.

alpha—improving annualized returns by 2–3%—they remain bound by the information contained in the tabular data.

The introduction of LLM-based architectures marks a structural break in predictive performance. Our framework, particularly when instantiated with GPT-4o, achieves a dominant out-of-sample R^2 of 2.76% and annualized returns exceeding 25.2% in the Top-1000 liquid subset. Crucially, DeepSeek-7B and GPT-4 variants significantly outperform their tabular counterparts (TabNet). This performance gap implies that the "modality gap" is a binding constraint: LLMs succeed not merely because they are better regressors, but because they leverage semantic reasoning to extract latent signals—such as sentiment shifts or thematic risks—that are invisible to pure numerical models. The results validate our hypothesis that combining semantic intelligence with numerical rigor yields a unique source of alpha.

C. Ablation Study

To disentangle the sources of performance gains, we perform a component-wise dismantling of the framework, as detailed in Table II. A central question is whether LLMs engage in semantic reasoning or function as over-parameterized regressors. To resolve this, we implement a "Semantic-Agnostic" ablation ($-$Semantics) where specific feature identifiers are replaced with generic tokens (e.g., "Feature_A"), while preserving numerical values. As shown in the table, this semantic masking precipitates a significant deterioration in Sharpe ratio from 1.63 to 1.38. This empirical gap isolates the value of linguistic comprehension, confirming that the LLM leverages pre-trained economic knowledge to contextualize numerical inputs.

TABLE II

ABLATION STUDY RESULTS. CS: CAUSAL SELECTION, RA: REGIME ADAPTIVITY, LLM: LANGUAGE MODEL, CH: CALIBRATION HEAD, DW: DYNAMIC WEIGHTING, NS: NAIVE SERIALIZATION. **NOTE: '$-$*Semantics*' USES ANONYMOUS TOKENS TO ISOLATE THE VALUE OF SEMANTIC REASONING.**

Variant	R^2%	RMSE	Ann. Ret.	Sharpe
Full model (ours)	**1.72**	0.075	22.7	1.63
$-$ CS (all factors)	1.48	0.082	18.4	1.29
$-$ RA (fixed wts.)	1.36	0.086	17.1	1.18
$-$ Semantics (Anon. Prompt)	1.51	0.079	19.5	1.38
$-$ LLM (numeric NN)	1.21	0.092	15.8	1.07
$-$ CH (removed)	1.43	0.081	16.7	1.23
$-$ DW $\rightarrow$ static avg.	1.29	0.087	16.2	1.14
$-$ NS (concat.)	1.18	0.095	15.1	1.02

Beyond semantic reasoning, the structural integrity of the pipeline is vital. Removing the causal discovery module causes a sharp degradation in the Sharpe ratio despite a moderate drop in R^2, indicating that non-causal factors fail to generalize and increase downside volatility out-of-sample. Causal screening acts as a "safety valve" to reject spurious signals that amplify tail risk. Similarly, the necessity of regime adaptation is confirmed by the performance decline when replacing the dynamic gating network with static weights. This static approach fails during the post-2020 inflationary period, illustrating that factor efficacy requires a model capable of pivoting as market narratives shift from "growth-driven" to "rate-driven".

The ablation results finally underscore the *calibration imperative*. Without the Calibration Head, the raw LLM exhibits "arithmetic hallucinations"—correctly identifying the direction of returns but failing to estimate magnitude accurately, which leads to suboptimal portfolio weighting. The calibration head aligns the semantic confidence of the LLM with the rigorous distribution of financial returns, proving indispensable for converting reasoning into tradable signals. Rolling-window evaluations corroborate these findings, revealing that the synergy between causal selection, regime adaptation, and calibrated semantic reasoning is structural rather than merely additive.

D. OOD Generalization: Chinese Credit Bond Market

TABLE III
OUT-OF-SAMPLE PREDICTION PERFORMANCE ON MONTHLY CREDIT BOND RETURNS IN CHINA (2013–2023). WE REPORT R^2%, RMSE, ANNUAL RETURN (%), AND SHARPE RATIO.

Model	R^2%	RMSE	Ann. Ret.	Sharpe
OLS	-1.92	0.089	4.7	0.39
PCR	0.37	0.084	6.9	0.56
RF	0.81	0.080	8.5	0.72
NN4	1.02	0.078	9.3	0.81
TabNet	1.41	0.073	11.7	1.06
DeepSeek-7B	1.63	0.071	13.6	1.23
GPT-4	1.86	0.069	15.2	1.39
GPT-4o	**2.14**	**0.066**	**17.1**	**1.57**

To test the framework's generalizability beyond developed equities, we extend the analysis to the Chinese Corporate Bond market (2013–2024), a venue characterized by lower efficiency, higher retail participation, and distinct regulatory regimes. As reported in Table III, our model demonstrates remarkable transferability. It achieves an R^2 of 2.14% and an annualized return of 17.1%, significantly outpacing local baselines. This result suggests that the "Causal-Regime-LLM" architecture captures universal properties of financial price formation—structural shifting and information asymmetry—rather than overfitting to specific characteristics of the U.S. equity market.

V. CONCLUSION

In this work, we demonstrate that the next frontier of financial intelligence relies on disciplined structural constraints rather than unconstrained parameter scaling. By synthesizing causal inference, continuous regime adaptation, and calibrated large language models, our proposed Neuro-symbolic architecture effectively resolves the trilemma of high-dimensional noise, non-stationarity, and modality misalignment. Empirical validation confirms that the causal information bottleneck acts as a critical regularizer against overfitting, while the calibration mechanism grounds semantic reasoning in statistical reality. Ultimately, this "structure-first" paradigm provides a robust blueprint for generalizable and trustworthy AI in finance, paving the way for future explorations into high-frequency microstructure adaptation and explainable agentic execution.

ACKNOWLEDGMENT

This work was supported in part by the National Key Research and Development Program of China under Grant 2023YFC3305402 and in part by the National Natural Science Foundation of China under Grant 62172053 and Grant 62302059.

REFERENCES

[1] Feng G, Giglio S, Xiu D. Taming the factor zoo: A test of new factors[J]. The Journal of Finance, 2020, 75(3): 1327-1370.
[2] Harvey C R, Liu Y. A census of the factor zoo[J]. Available at SSRN 3341728, 2019.
[3] Ahuja K, Shanmugam K, Varshney K, et al. Invariant risk minimization games[C]//International Conference on Machine Learning. PMLR, 2020: 145-155.
[4] Chib S, Smith S. Factor Selection and Structural Breaks[J]. 2024.
[5] Andersen T G, Varneskov R T. Testing for parameter instability and structural change in persistent predictive regressions[J]. Journal of Econometrics, 2022, 231(2): 361-386.
[6] Loughran T, McDonald B. Textual analysis in finance[J]. Annual Review of Financial Economics, 2020, 12(1): 357-375.
[7] Cohen L, Malloy C, Nguyen Q. Lazy prices[J]. The Journal of Finance, 2020, 75(3): 1371-1415.
[8] Bhatia G, Cavusoglu H, Abdul-Mageed M. Fintral: A family of gpt-4 level multimodal financial large language models[C]//Findings of the Association for Computational Linguistics: ACL 2024. 2024: 13064-13087.
[9] Gruver N, Finzi M, Qiu S, et al. Large language models are zero-shot time series forecasters[J]. Advances in Neural Information Processing Systems, 2023, 36: 19622-19635.
[10] Shen Y, Shao J, Zhang X, et al. Large language models empowered autonomous edge AI for connected intelligence[J]. IEEE Communications Magazine, 2024, 62(10): 140-146.
[11] Huang B, Zhang K, Gong M, et al. Causal discovery and forecasting in nonstationary environments with state-space models[C]//International conference on machine learning. Pmlr, 2019: 2901-2910.
[12] Cekic S, Grandjean D, Renaud O. Time, frequency, and time-varying Granger-causality measures in neuroscience[J]. Statistics in medicine, 2018, 37(11): 1910-1931.
[13] Pamfil R, Sriwattanaworachai N, Desai S, et al. Dynotears: Structure learning from time-series data[C]//International Conference on Artificial Intelligence and Statistics. Pmlr, 2020: 1595-1605.
[14] Huang B, Zhang K, Zhang J, et al. Causal discovery from heterogeneous/nonstationary data[J]. Journal of Machine Learning Research, 2020, 21(89): 1-53.
[15] Gorishniy Y, Rubachev I, Khrulkov V, et al. Revisiting deep learning models for tabular data[J]. Advances in neural information processing systems, 2021, 34: 18932-18943.
[16] Zeng A, Chen M, Zhang L, et al. Are transformers effective for time series forecasting?[C]//Proceedings of the AAAI conference on artificial intelligence. 2023, 37(9): 11121-11128.
[17] Cai Z, Chang S Y. A new test on asset return predictability with structural breaks[J]. Journal of Financial Econometrics, 2024, 22(4): 1042-1074.
[18] Zhu F, Liu Z, Feng F, et al. Tat-llm: A specialized language model for discrete reasoning over financial tabular and textual data[C]//Proceedings of the 5th ACM International Conference on AI in Finance. 2024: 310-318.
[19] Zhu F, Li J, Pan L, et al. FinTMMBench: Benchmarking Temporal-Aware Multi-Modal RAG in Finance[J]. arXiv preprint arXiv:2503.05185, 2025.
[20] Bonnier T. Revisiting multimodal transformers for tabular data with text fields[C]//Findings of the Association for Computational Linguistics ACL 2024. 2024: 1481-1500.
[21] Zheng X, Dan C, Aragam B, et al. Learning sparse nonparametric dags[C]//International conference on artificial intelligence and statistics. Pmlr, 2020: 3414-3425.
[22] Leyder S, Raymaekers J, Verdonck T. TSLiNGAM: DirectLiNGAM under heavy tails[J]. Journal of Computational and Graphical Statistics, 2025, 34(2): 437-447.
[23] Runge J, Nowack P, Kretschmer M, et al. Detecting and quantifying causal associations in large nonlinear time series datasets[J]. Science advances, 2019, 5(11): eaau4996.
[24] Arik S Ö, Pfister T. Tabnet: Attentive interpretable tabular learning[C]//Proceedings of the AAAI conference on artificial intelligence. 2021, 35(8): 6679-6687.
[25] Huang X, Khetan A, Cvitkovic M, et al. Tabtransformer: Tabular data modeling using contextual embeddings[J]. arXiv preprint arXiv:2012.06678, 2020.