AMES INVESTMENT SYSTEMS RESEARCH

Regime-Conditioned Information Routing in Institutional Quantitative Trading

A Formal Theory, Evidence Synthesis, and Falsifiable Validation Standard

Ogdn Ames

Ames Investment Systems

September 21, 2026

Author note. This paper is a technical research synthesis and theoretical framework. It does not report an audited live investment record and does not constitute investment advice. Exact securities, vendor mappings, production feature definitions, architecture dimensions, execution rules, and deployment thresholds are intentionally omitted. Empirical results attributed to external studies are reported as study-specific evidence, not as guarantees of replicable or live performance.

Correspondence: Ames Investment Systems.

Abstract

Transformer architectures are increasingly used in financial forecasting and systematic trading, but the relevant scientific question is not whether attention is expressive. It is whether adaptive information routing earns its statistical and economic complexity under realistic institutional constraints. This paper develops *Regime-Conditioned Information Routing* (RCIR), a falsifiable theory of when transformer-like routing should add value in quantitative finance. The theory distinguishes point-in-time information availability from semantic content, models predictive relevance as state-dependent, and treats attention as a routing operator rather than an autonomous source of alpha. In a stylized conditional-linear environment, we show that an oracle regime-adaptive predictor weakly dominates the best static linear predictor and that the exact population risk gap equals a covariance-weighted dispersion of regime-specific predictive coefficients. The result identifies the source of any potential routing advantage: variation in the predictive mapping itself, not architectural novelty. We then derive practical implications for finite samples, where routing gains must exceed estimation error, model-selection bias, turnover, market impact, and governance costs.

The paper synthesizes finance-specific and general time-series evidence through September 21, 2026. Recent results support three conclusions. First, structured hybrid models and variable-selection mechanisms can outperform plain transformer variants in finance-specific benchmarks. Second, finance-specific foundation models such as FinCast and Kronos strengthen representation-learning evidence, but forecasting improvements do not by themselves establish persistent tradable alpha. Third, recent financial-return benchmarking shows that pretrained time-series foundation models can rank well while producing only sparse statistically reliable improvements over naive return benchmarks. These findings are consistent with RCIR: attention is most defensible when predictive relevance is sparse, changes across states, spans multiple horizons, or depends on structured cross-asset context; it should add little when a stable low-dimensional sufficient representation exists.

To make the theory empirically refutable, we define a target-free Regime Reallocation Index, five prespecified hypotheses, an architecture-neutral model ladder, and a frozen forward-validation protocol with point-in-time data, purged rolling evaluation, multiple-comparison controls, realistic costs, and explicit promotion gates. The resulting research standard treats transformers as optional adaptive routing components inside a disciplined forecasting, portfolio-construction, and execution process. The central claim is deliberately narrow: a transformer deserves deployment only when it demonstrates incremental net out-of-sample value precisely where state-dependent information routing predicts that it should.

Keywords: transformers; quantitative trading; financial time series; regime change; attention; foundation models; systematic macro; model selection; backtest overfitting; point-in-time data.

Contents

1 Introduction **1**

2 Research Question, Scope, and Claim Discipline **2**
2.1 The research question 2
2.2 Three distinct claim levels 2
2.3 Evidence grades are claim-specific 2

3 Why Financial Sequence Modeling Is a Different Problem **3**
3.1 The filtration is part of the model 3
3.2 Returns are a hostile target 3
3.3 The target is a decision, not a generic forecast 4

4 Regime-Conditioned Information Routing **4**

4.1 Core intuition . . . 4
4.2 Formal setup . . . 5
4.3 State-dependent temporal memory . . . 6
4.4 Variable selection as an information bottleneck . . . 6
4.5 Cross-asset attention requires structural friction . . . 6
4.6 Pretraining is a prior, not a source of information . . . 6

5 Finite-Sample Economics: When Population Routing Value Is Not Enough 7
5.1 The complexity burden rises as data scarcity increases . . . 7
5.2 Attention weights are diagnostics, not causal explanations . . . 7

6 Operationalizing Regime Change Without Look-Ahead Bias 8

7 Falsifiable Predictions of RCIR 8

8 A Stronger Empirical Design 9
8.1 Architecture-neutral model ladder . . . 9
8.2 Point-in-time dataset contract . . . 10
8.3 Chronological evaluation . . . 10
8.4 Frozen model-selection budget . . . 10
8.5 Forecast metrics and statistical tests . . . 10
8.6 Cost and capacity stress . . . 11
8.7 The primary RCIR test . . . 11

9 What the 2021–2026 Evidence Actually Supports 11

10 Foundation Models: What Changed After 2025 13

11 Institutional Validation and Governance Standard 13
11.1 Minimum promotion gates . . . 13
11.2 Research ledger and model cards . . . 14
11.3 Online adaptation deserves special skepticism . . . 14

12 Interpretability and Failure Analysis 14
12.1 Interpretability should answer intervention questions . . . 14
12.2 Dominant failure modes . . . 14

13 Research Architecture Without Revealing a Proprietary Strategy 15

14 Discussion 15

15 Limitations and Non-Claims 16

16 Conclusion 16

A Proofs 17
A.1 Proof of Proposition 1 . . . 17
A.2 Closed-form static coefficient . . . 17
A.3 Interpretation under equal covariance . . . 18

B Preregistered RCIR Experiment Template **18**

C Non-Proprietary Institutional Research Checklist **18**

D Claim Boundary Matrix **19**

1. Introduction

The strongest version of the transformer-in-finance claim is usually the wrong one. A transformer is not an economic anomaly, and an architecture specification is not evidence of persistent excess return. The institutional question is narrower and more demanding: *when does adaptive routing of already-available information improve a decision after estimation error, data leakage controls, transaction costs, and model-selection bias are accounted for?*

That question matters because financial data differ materially from the domains in which transformers first became dominant. Natural-language corpora contain enormous amounts of repeated structure. Institutional trading often offers one realized market history, weak and unstable signals, asynchronous releases, revised macroeconomic series, changing index membership, endogenous transaction costs, and targets only indirectly related to generic forecast error. A model can rank well on mean-squared error and still be economically useless after turnover, spread, financing, impact, position limits, and selection bias. Conversely, a modest predictive improvement can matter when it is stable, diversifying, implementable, and concentrated in economically important states.

The literature now supports neither blanket enthusiasm nor blanket dismissal. Temporal Fusion Transformers (TFT) showed that variable selection, gating, local recurrent processing, and attention can be combined effectively in multi-horizon forecasting (Lim et al., 2021). The Momentum Transformer applied attention-based temporal adaptation to systematic trading (Wood et al., 2021). More recent finance-specific evidence complicates the picture. A 2026 large-scale financial benchmark reports that structured recurrent and hybrid models outperform several generic transformer variants on a common futures task, with a Variable-Selection-Network-plus-LSTM family leading the primary risk-adjusted ranking (Saly-Kaufmann et al., 2026). DeePM further reports that constrained cross-asset interaction, lag discipline, robust objectives, and explicit costs matter more than unconstrained attention alone (Wood et al., 2026). At the same time, FinCast and Kronos provide stronger evidence that domain-specific financial pretraining can improve representation and forecasting tasks (Zhu et al., 2025; Shi et al., 2026). Yet a 2026 study of pretrained time-series foundation models on U.S. equity returns finds that strong model rankings coexist with sparse statistically reliable gains over random-walk baselines (Noguer I Alonso and Pereira Franklin, 2026).

This paper proposes Regime-Conditioned Information Routing (RCIR) as a way to reconcile those findings. RCIR is not a claim that attention creates information. It is a claim about *allocation*: when the predictive value of variables, lags, events, or cross-asset relationships changes with market state, an adaptive router can reduce misspecification relative to a fixed mapping. If the relevant mapping is stable, low-dimensional, and approximately linear, the router has little population advantage and may lose in finite samples.

The paper makes five contributions. First, it formalizes RCIR in a stylized model and derives an exact population risk gap between static and regime-adaptive predictors. Second, it converts the qualitative idea of "regime change" into a target-free Regime Reallocation Index (RRI) that can be computed using information available at the decision time. Third, it separates representational evidence, forecasting evidence, and trading evidence, preventing a forecasting result from being silently promoted into an alpha claim. Fourth, it specifies falsification tests, not just supportive tests. Fifth, it proposes an institutional validation ladder in which complexity must earn promotion through frozen, cost-aware, selection-adjusted evidence.

The resulting thesis is intentionally conservative: transformers are credible in institutional quantitative research when they operate as constrained, point-in-time information routers inside a broader process. They should not be assumed necessary, and they should not be credited for performance that can be reproduced by simpler models on the same information set.

2. Research Question, Scope, and Claim Discipline

2.1. The research question

Let $Y_{t,h}$ denote a forecasting or decision target at horizon h, and let $\mathcal{F}_t^{\text{PIT}}$ denote the complete information filtration that was genuinely available immediately before the decision at time t. The practical research question is

$$A_t = A(\mathcal{F}_t^{\text{PIT}}) \quad \Longrightarrow \quad \text{incremental net out-of-sample decision value?} \tag{1}$$

The phrase "net out-of-sample" is essential. A routing architecture that improves in-sample fit but raises estimation variance, turnover, or selection bias has not earned its complexity.

2.2. Three distinct claim levels

Financial machine-learning papers often collapse three different propositions. RCIR keeps them separate:

1. **Representation claim:** a model learns reusable temporal or cross-sectional structure.
2. **Forecasting claim:** a model improves a prespecified predictive loss on future data.
3. **Trading claim:** a model improves an economically relevant objective after costs, constraints, and research selection.

A model can satisfy the first without the second, or the second without the third. This is especially important for foundation models. Chronos demonstrates that large-scale pretraining can support competitive zero-shot time-series forecasting across heterogeneous domains (Ansari et al., 2024). FinCast and Kronos extend that idea toward financial data (Zhu et al., 2025; Shi et al., 2026). None of those results, by itself, is proof of persistent live alpha.

2.3. Evidence grades are claim-specific

Table 1 defines the evidence standard used throughout this paper. A paper can receive a high grade for architecture or forecasting and a lower grade for live trading. Evidence grades apply to claims, not to entire papers.

Table 1: Claim-specific evidence grading

Grade	Minimum standard	Typical examples	Permitted interpretation
A	Multiple independent literatures or highly credible replicated evidence; point-in-time design where relevant	Real-time macro vintage evidence; established forecast-comparison methods; replicated statistical regularities	Strong methodological or empirical support; still not a guarantee of future returns
B	Repeated out-of-sample evidence with meaningful external validation, ablations, or realistic constraints	Finance-specific benchmarks with chronological evaluation, costs, multiple seeds, or strong robustness analysis	Reasonable research evidence; deployment still requires independent validation
C	Technically plausible engineering evidence with limited direct trading validation	General time-series architecture improvements; finance forecasting without portfolio implementation	Supports mechanism or representation claim only
D	Speculative or emerging claim with insufficient independent evidence	Autonomous alpha claims inferred from generic model capability; single backtest without selection controls	Research hypothesis, not an investment claim

3. Why Financial Sequence Modeling Is a Different Problem

3.1. The filtration is part of the model

The first source of bias in financial machine learning is often not the architecture. It is the data pipeline. Macroeconomic series are revised; consensus estimates change before releases; index membership changes; corporate filings arrive after fiscal periods end; and some vendor histories overwrite old values with revised values. Croushore and Stark's real-time macroeconomic data work shows directly why vintage information can change historical forecast evaluation (Croushore and Stark, 2001). A modern model trained on today's revised history can therefore consume information that was not available to a historical decision maker.

RCIR treats information availability as a first-class state variable. For observation x_j describing economic period τ_j, define

$$u_j = \text{Encoder}(\text{value}_j, \text{type}_j, \tau_j, a_j, v_j, m_j), \tag{2}$$

where a_j is the release or availability time, v_j is the data-vintage or revision identity, and m_j contains metadata. The model may condition on token j at decision time t only if $a_j \le t$.

This distinction is not cosmetic. Semantic time and information time are different coordinates. A March payroll observation and a March equity return may refer to the same calendar period while entering the investable information set on different dates. Forward-filling the payroll release across daily rows without preserving the release event can destroy that distinction and create artificial temporal regularity.

3.2. Returns are a hostile target

Financial returns have weak signal-to-noise ratios, nonstationarity, heavy tails, serial dependence in strategy returns, and a single realized history. These features make flexible models unusually vulnerable to data mining. White's Reality Check formalizes the danger of repeatedly searching one historical record (White, 2000). The Deflated Sharpe Ratio adjusts performance inference for multiple testing and non-normality (Bailey and López de Prado, 2014); the Probability of Backtest Overfitting provides a complementary framework for estimating selection risk in strategy searches (Bailey et al., 2016). Lo shows why naive Sharpe annualization can be misleading under

serial correlation (Lo, 2002).

The implication is structural: a financial transformer must be evaluated as one candidate inside a research process, not as an isolated forecast model.

3.3. The target is a decision, not a generic forecast

Let a forecast model produce a predictive distribution

$$p_\theta(Y_{t,h} \mid \mathcal{F}_t^{\mathrm{PIT}}). \tag{3}$$

A robust institutional design should usually keep this representation stage separate from portfolio construction. The downstream decision may solve

$$w_t^* = \arg\max_{w \in \mathcal{W}} \left\{ \mathbb{E}_t[w^\top r_{t+1}] - \lambda_R \mathcal{R}(w) - \lambda_T \mathcal{T}(w, w_{t-1}) - \lambda_I \mathcal{I}(w) \right\}, \tag{4}$$

where $\mathcal{R}$ is risk, $\mathcal{T}$ turnover cost, and $\mathcal{I}$ impact or implementation cost. Forecast quality and economic utility are related, but not identical.

4. Regime-Conditioned Information Routing

4.1. Core intuition

Attention is economically useful only when the mapping from information to expected return, risk, or another decision target is conditional on state. If the optimal predictive relation is stable and low-dimensional, a transformer is an inefficient estimator. If the relevant variable set, temporal lag, event type, or cross-asset channel changes over time, a static model incurs misspecification.

RCIR treats a transformer-style module as a learned routing operator. At decision time t, the operator allocates representation capacity across causally available tokens conditional on an inferred state:

$$A_t = \mathrm{Route}_\theta(U_{\le t}, \hat{Z}_t, h), \tag{5}$$

where $U_{\le t}$ is the point-in-time token set and $\hat{Z}_t$ is a state estimate produced only from $\mathcal{F}_t^{\mathrm{PIT}}$. The forecasting head then computes

$$\widehat{Y}_{t,h} = g_\theta(A_t, U_{\le t}). \tag{6}$$

The claim is not that A_t creates information. The claim is that a flexible router can reduce the error caused by using the wrong information subset or the wrong temporal/cross-sectional weighting in a changing environment.

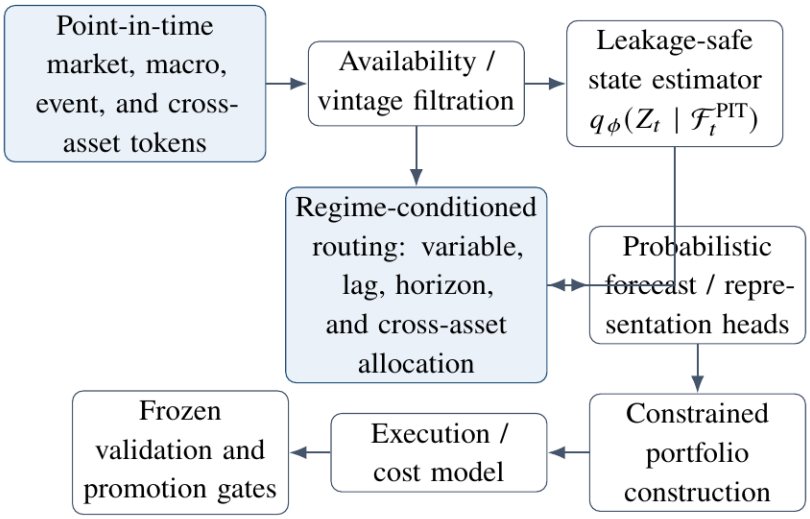

Figure 1: RCIR architecture as a research stack. Attention is one routing mechanism inside a point-in-time and economically constrained process.

4.2. Formal setup

Consider a stylized one-step target Y_t and feature vector $X_t \in \mathbb{R}^p$. Let $Z_t \in \{1, \ldots, K\}$ denote an economic state. The state is not assumed to be directly observable; the oracle analysis below asks what value state-conditioned mapping could have in population before accounting for estimation error.

Assumption 1 (Conditional linear signal)**.** *For each state z,*

$$Y_t = \beta_z^\top X_t + \varepsilon_t, \qquad \mathbb{E}[\varepsilon_t \mid X_t, Z_t = z] = 0, \tag{7}$$

with finite second moments and state-conditional feature covariance $\Sigma_z = \mathbb{E}[X_t X_t^\top \mid Z_t = z]$ positive semidefinite.

The best state-adaptive oracle predictor is $f_A(X_t, Z_t) = \beta_{Z_t}^\top X_t$. The best static linear predictor uses a single coefficient vector

$$\beta^* = \arg\min_{b \in \mathbb{R}^p} \mathbb{E}[(Y_t - b^\top X_t)^2]. \tag{8}$$

Proposition 1 (Population value of regime-conditioned routing)**.** *Under Assumption 1, the excess mean-squared prediction risk of the best static linear predictor relative to the oracle regime-adaptive predictor is*

$$\boxed{\mathcal{G}_{\text{route}} = \mathbb{E}\left[(\beta_{Z_t} - \beta^*)^\top \Sigma_{Z_t} (\beta_{Z_t} - \beta^*)\right] \geq 0.} \tag{9}$$

Equality holds if and only if $\beta_{Z_t} = \beta^*$ on the covariance-supported feature subspace almost surely.

The proof is given in Appendix A. Equation (9) is the central theoretical statement of RCIR. It identifies the population source of routing value as a covariance-weighted dispersion in the predictive mapping. The result does not mention transformers. Any estimator capable of learning the state-conditioned mapping can capture the gain. A transformer is therefore justified only if its routing mechanism estimates that mapping more effectively than alternatives.

Corollary 1 (Stable sufficient mapping)**.** *If $\beta_z = \bar{\beta}$ for all economically relevant states z, then $\mathcal{G}_{\text{route}} = 0$. In the population, no state-conditioned routing architecture can improve squared-error prediction solely by changing weights across states.*

This corollary formalizes a practical falsifier: if a stable low-dimensional model performs equally well after costs

and proper validation, RCIR predicts that the transformer has not earned its complexity.

4.3. State-dependent temporal memory

The same logic applies to lags. Let $X_{t-\ell}$ denote a candidate lagged token. A generic attention score can be written

$$\alpha_{t,\ell,j} = \text{Normalize}\left(s_\theta(q_t, k_{t-\ell,j}, \hat{Z}_t, h)\right), \tag{10}$$

where j indexes variable or event type. RCIR requires the score to be capable of changing with $\hat{Z}_t$ and horizon h. The economic hypothesis is not that distant lags always matter. It is that the useful half-life of information is state-dependent. Monetary-policy evidence provides an intuitive example: policy actions and forward-path communication can affect different parts of the yield curve differently (Gurkaynak et al., 2005). A fixed memory kernel may therefore be misspecified when the relevant channel changes.

4.4. Variable selection as an information bottleneck

Financial panels often contain hundreds of correlated transforms of a small number of economic dimensions. In this environment, unrestricted attention can spend capacity rediscovering redundancy. Variable-selection networks, as used in TFT, act as a learned bottleneck (Lim et al., 2021). The 2026 finance benchmark's strong performance for variable-selection-plus-recurrent hybrids is consistent with the idea that compression and temporal state encoding can matter more than generic attention (Saly-Kaufmann et al., 2026).

Define a gating vector $g_t \in [0,1]^p$ and routed representation

$$\tilde{X}_t = g_t \odot X_t, \qquad g_t = g_\theta(X_{\le t}, \hat{Z}_t). \tag{11}$$

RCIR predicts that dynamic selection should add the most when the raw panel is redundant and the identity of the informative subset changes through time. If selection weights are nearly constant and a fixed sparse subset performs equally well, the dynamic router is unnecessary.

4.5. Cross-asset attention requires structural friction

A flexible cross-asset attention layer can fit thousands of unstable pairwise relations. That is dangerous when the number of effective regimes is small relative to the number of possible interactions. DeePM reports that unrestricted cross-attention underperformed more structured alternatives in its macro setting, while graph regularization improved results (Wood et al., 2026). RCIR interprets this not as evidence against cross-asset learning, but as evidence that economically admissible interaction spaces should be constrained.

One generic formulation is

$$\alpha^{\text{cross}}_{i,j,t} \propto \exp(s_{i,j,t})\, M_{i,j,t}, \tag{12}$$

where $M_{i,j,t} \in [0,1]$ is an admissibility mask or shrinkage prior based on lag discipline, asset class, economic linkage, or a predeclared graph. The mask need not be literally binary. Its purpose is to make unsupported interaction expensive.

4.6. Pretraining is a prior, not a source of information

A pretrained time-series model imports regularities learned from a broader corpus. Chronos shows the general feasibility of large-scale cross-domain pretraining (Ansari et al., 2024); Time-MoE scales that approach to billion-parameter models (Shi et al., 2025). FinCast and Kronos provide finance-specific versions of the idea

(Zhu et al., 2025; Shi et al., 2026). RCIR interprets pretraining as *prior compression*: it can reduce the amount of target-domain data needed to learn useful representations, but it cannot increase the information-theoretic predictability of the target.

This distinction explains why pretrained models can improve average rankings while barely outperforming naive return benchmarks in statistically significant terms. Noguer I Alonso and Pereira Franklin report exactly that pattern for U.S. equity returns (Noguer I Alonso and Pereira Franklin, 2026). Representation quality and tradable return predictability are different objects.

5. Finite-Sample Economics: When Population Routing Value Is Not Enough

Proposition 1 is an oracle result. Institutional research operates in finite samples. Let $\widehat{f}_A$ be a learned adaptive model and $\widehat{f}_S$ a learned static baseline. A useful decomposition is

$$\Delta_{\text{net}} \approx \mathcal{G}_{\text{route}} - C_{\text{est}} - C_{\text{select}} - C_{\text{turn}} - C_{\text{impact}} - C_{\text{ops}}, \tag{13}$$

where C_{est} is additional estimation error from flexibility, C_{select} is research-selection inflation, C_{turn} and C_{impact} are trading frictions, and C_{ops} captures latency, compute, monitoring, and governance burden.

Equation (13) is not a theorem; it is an institutional accounting identity. It states the burden of proof. A flexible architecture deserves capital only when observed routing gains are larger than the costs created by obtaining them.

5.1. The complexity burden rises as data scarcity increases

In language modeling, additional parameters can be supported by massive corpora. In finance, a strategy may have only a few dozen economically distinct regimes. The effective sample size can be much lower than the number of timestamps because observations share macro conditions, volatility states, and overlapping labels. Consequently, parameter count is a weak proxy for complexity; what matters is the number of independently informative situations relative to the model's degrees of freedom.

This leads to a practical research principle:

$$\boxed{\text{The burden of proof should rise with effective model flexibility.}} \tag{14}$$

A 50-million-parameter model that improves gross Sharpe by a few basis points over a tuned linear model has not necessarily contributed useful information. A structured model that survives frozen forward testing, cost stress, multiple seeds, and regime-specific falsification is more credible even if its headline backtest is lower.

5.2. Attention weights are diagnostics, not causal explanations

An attention heatmap can show where a network allocates representation mass. It does not establish that the attended token caused the forecast or the realized return. Better tests include ablation, counterfactual delay, feature-group removal, structured perturbation, and representation probing. A claimed information channel should alter the model's behavior in a stable and economically coherent direction when the channel is removed or delayed.

6. Operationalizing Regime Change Without Look-Ahead Bias

The word "regime" is often used loosely enough to become unfalsifiable. RCIR therefore separates the latent economic state from an observable, target-free proxy for information reallocation.

Definition 1 (Regime Reallocation Index)**.** *Let W_t be a rolling window ending at decision time t. Define a vector of distribution-shift statistics $d_t = (d_t^{\text{cov}}, d_t^{\text{corr}}, d_t^{\text{scale}}, d_t^{\text{breadth}}, d_t^{\text{event}})$ computed only from information in $\mathcal{F}_t^{\text{PIT}}$. The Regime Reallocation Index is*

$$\text{RRI}_t = \sum_{m=1}^{M} \omega_m \, \text{rz}_t(d_t^{(m)}), \qquad \omega_m \geq 0, \quad \sum_m \omega_m = 1, \tag{15}$$

where $\text{rz}_t(\cdot)$ is a robust expanding-window standardization and weights ω_m are fixed using training data only.

A concrete implementation can use the components in Table 2. The exact specification should be preregistered and frozen before the final test period.

Table 2: Example target-free components of the Regime Reallocation Index

Component	Example statistic	Economic interpretation
Covariance shift	$\|\widehat{\Sigma}_t - \widehat{\Sigma}_{t-L}\|_F / (\|\widehat{\Sigma}_{t-L}\|_F + \epsilon)$	Change in multivariate scale and dependence structure
Correlation shift	$\|\widehat{C}_t - \widehat{C}_{t-L}\|_F$	Rewiring of cross-asset co-movement independent of marginal volatility
Scale shift	Median absolute change in rolling realized volatility across assets	Change in noise level and risk transmission
Breadth / concentration	Change in leading covariance eigenvalue share or participation ratio	Change in whether one common factor or many dimensions dominate
Event intensity	Prespecified count or weighted severity of newly arriving scheduled macro/event tokens	Change in arrival rate of discrete information

Two design constraints matter. First, RRI_t cannot use future returns. Second, any thresholds used to define "high" and "low" reallocation must be estimated using training history only. Retrospective crisis labels are useful for description but weak evidence for a forward-deployable regime mechanism.

7. Falsifiable Predictions of RCIR

A useful theory should be capable of losing. Table 3 states five preregisterable predictions.

Table 3: RCIR hypotheses and falsification tests

ID	Prediction	Primary test	Falsification signal
H1	Routing advantage is concentrated when information relevance is changing	Test adaptive-minus-static loss or utility by frozen RRI quantile	No monotonic uplift; gain concentrated in low-RRI states
H2	Dynamic variable selection matters most in redundant panels	Compare dynamic gate, fixed selected subset, and no-selection model with downstream stack held constant	Dynamic selection fails to improve high-redundancy subsets
H3	Structured cross-asset interaction is more stable than unconstrained attention when interaction dimensionality is large	Graph/mask ablation; compare seed dispersion, forward loss, and turnover	Unconstrained attention is equally or more stable with no cost penalty
H4	Domain-specific pretraining helps most in low-data and representation tasks	Frozen-sample-size transfer experiment across raw returns, volatility, and multi-series representation targets	Pretraining advantage grows with abundant target data or only on near-noise raw returns
H5	Transformer advantage vanishes when a stable low-dimensional sufficient state representation is available	Compare RCIR model to compressed linear/SSM baseline using the same filtered information	Persistent net advantage remains after stable sufficient compression with no state dependence

H1 is the distinctive empirical prediction. A transformer that merely has a higher average score does not confirm RCIR. The incremental gain should appear where the theory says adaptive routing is valuable.

8. A Stronger Empirical Design

8.1. Architecture-neutral model ladder

The empirical test should not be framed as "transformer versus old model." It should be a controlled sequence in which each layer of flexibility must earn itself. Table 4 gives a minimum ladder.

Table 4: Minimum model ladder for an RCIR test

Model	Class	Purpose	Key control
M_0	Linear / distributed lag / ridge	Tests whether the task is basically low-dimensional and stable	Same point-in-time features and target
M_1	Tree ensemble or sparse nonlinear model	Captures threshold and interaction effects without sequence attention	Same information set; tuned chronologically
M_2	LSTM, xLSTM, or state-space model	Tests whether adaptive memory alone explains gains	Comparable training protocol and cost treatment
M_3	Plain transformer	Measures value of generic self-attention	No privileged features or additional data
M_4	RCIR-structured model	Adds dynamic selection, state-conditioned routing, and constrained cross-asset interaction	Ablate one component at a time

Recent evidence makes this ladder necessary. Zeng et al. show that simple linear models can beat sophisticated transformers on standard long-horizon forecasting datasets (Zeng et al., 2023). PatchTST demonstrates that tokenization and channel design can materially alter time-series transformer performance (Nie et al., 2023). iTransformer shows that reversing the tokenization axis can improve multivariate forecasting (Liu et al., 2024). Finance-specific benchmarking then shows that these generic rankings do not transfer automatically to economic objectives (Saly-Kaufmann et al., 2026). The model class is therefore not the hypothesis; the conditional information problem is.

8.2. Point-in-time dataset contract

Every observation should carry at least four timestamps when relevant:

1. economic observation time;
2. public release time;
3. internal ingestion or knowledge time;
4. revision/vintage identity.

The research code should construct each historical decision row from the vintage that existed at that decision time. Corporate actions, index membership, delistings, futures rolls, and vendor corrections require similar treatment.

8.3. Chronological evaluation

Random train/test splits are inappropriate for most trading claims because they violate chronology and leak overlapping regimes. A serious design should use expanding or rolling training windows, a temporally later validation period, and one or more frozen test blocks. If labels overlap, purging and embargo should be applied around fold boundaries. Every transformation—scaling, PCA, feature selection, token normalization, missing-value rules, and state estimation—must be fit using training information only.

8.4. Frozen model-selection budget

The experiment count is part of the statistical evidence. For each serious trial, the ledger should record dataset snapshot, code commit, hypothesis, feature version, hyperparameters, seed, train/validation/test dates, cost model, and result. The final frozen test block should be consumed once. Repeated redesign after observing it converts "out-of-sample" into another training loop.

8.5. Forecast metrics and statistical tests

Forecasting evidence should report the prespecified primary loss and economically relevant secondary metrics. Pairwise forecast comparisons can use Diebold-Mariano tests when their assumptions are appropriate (Diebold and Mariano, 1995). When many models are compared, the Model Confidence Set provides a framework that acknowledges uncertainty in ranking (Hansen et al., 2011). Block bootstrap methods should preserve temporal dependence.

For trading outcomes, the paper should report at least net return, volatility, Sharpe with dependence-aware inference, drawdown, turnover, breakeven transaction cost, tail metrics, and seed dispersion. White's Reality Check, the Deflated Sharpe Ratio, or PBO-style selection analysis should be used when model search is material (White, 2000; Bailey and López de Prado, 2014; Bailey et al., 2016). No single statistic should carry the claim.

8.6. Cost and capacity stress

Gross backtests answer an incomplete question. The cost model should be applied at the same timestamp as the intended decision and include at minimum bid-ask or half-spread, commissions/fees where relevant, slippage assumptions, and a sensitivity analysis for impact. If the strategy trades multiple asset classes, costs should be asset-specific. A useful report includes the breakeven cost at which incremental utility versus the benchmark disappears.

8.7. The primary RCIR test

Let $L_{m,t}$ be the frozen out-of-sample loss of model m at time t. Define the adaptive routing uplift relative to a strong simple benchmark as

$$\Delta_t = L_{S,t} - L_{A,t}. \tag{16}$$

Partition the frozen test sample using RRI thresholds fixed from training data. The central prediction is

$$\mathbb{E}[\Delta_t \mid \mathrm{RRI}_t \in Q_4] > \mathbb{E}[\Delta_t \mid \mathrm{RRI}_t \in Q_1], \tag{17}$$

with analogous tests for net utility after the portfolio layer. A monotonic slope across RRI quantiles is stronger evidence than a single high-versus-low split. The same test should be repeated for each ablation. If the full model improves average performance but the gain has no relationship to information reallocation, the empirical result may still be useful, but it is not evidence for RCIR.

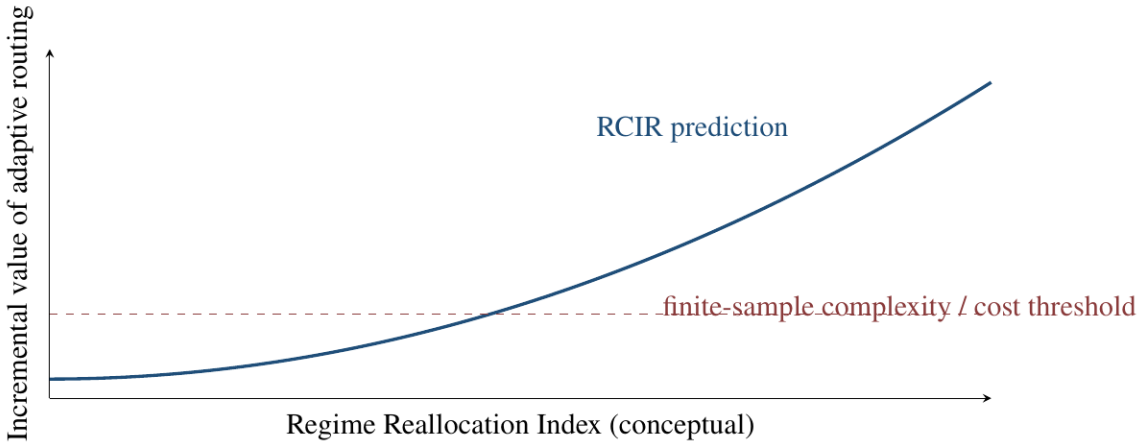

Figure 2: Conceptual RCIR prediction. This is not measured market data. Adaptive routing becomes economically attractive only when its state-dependent population gain exceeds finite-sample and implementation costs.

9. What the 2021–2026 Evidence Actually Supports

Table 5 summarizes the literature most relevant to the theory. The table deliberately distinguishes the result reported by each study from the inference that can reasonably be drawn for institutional trading.

Table 5: Selected evidence relevant to RCIR

Study	Task	Reported result	RCIR interpretation
Lim et al. (2021), TFT	General multi-horizon forecasting	Variable selection, gating, local recurrent processing, and interpretable self-attention improve performance on several real-world datasets	Supports conditional selection and hybrid inductive bias; not a trading result
Wood et al. (2021), Momentum Transformer	Futures trend / mean-reversion trading	Attention-LSTM hybrid reports improved performance including transaction-cost-aware results and adaptation around regime change	Finance-specific evidence that temporal routing can help; still study-specific backtest evidence
Zeng et al. (2023), DLinear	Long-horizon time-series forecasting	Simple linear models outperform several transformer variants on standard datasets	Strong warning that architectural complexity must defeat simple baselines
Nie et al. (2023), PatchTST	General forecasting	Patching and channel-independence improve transformer forecasting	Tokenization and representation design can dominate "attention versus no attention" framing
Liu et al. (2024), iTransformer	Multivariate forecasting	Variate tokens improve performance across standard datasets	Cross-variable structure can matter, but generic forecasting leadership is not equivalent to finance leadership
Ansari et al. (2024), Chronos	Foundation-model forecasting	Pretraining supports strong zero-shot forecasting across many datasets	Pretraining can supply a useful prior; does not create target-domain information
Zhu et al. (2025), FinCast	Financial time-series forecasting	Finance-oriented foundation model reports strong zero-shot forecasting across domains/resolutions	Stronger evidence for domain-specific representation transfer; trading implications remain indirect
Shi et al. (2025), Time-MoE	Large-scale TS foundation model	Sparse mixture-of-experts scaling improves general forecasting performance	Supports scalable routing/compression; finance-specific alpha remains untested
Shi et al. (2026), Kronos	Financial K-line forecasting, volatility, generation	Finance-specific tokenizer and pretraining on more than 12 billion K-line records; strong reported zero-shot benchmark gains	Domain-specific pretraining can learn reusable market representations; economic return claims remain separate
Saly-Kaufmann et al. (2026)	Financial forecasting and position sizing	Structured recurrent and hybrid models lead several risk-adjusted comparisons; generic transformer variants are not uniformly dominant	Direct support for task alignment, variable selection, and structured temporal modeling over architecture fashion
Wood et al. (2026), DeePM	Systematic macro portfolio management	Structured lagging, graph prior, robust objective, and explicit costs are important in reported ablations	Supports structural friction, causal timing discipline, and robust objectives around attention
Noguer I Alonso and Pereira Franklin (2026)	U.S. equity return forecasting	Pretrained TSFMs rank strongly, but statistically reliable gains over random-walk benchmarks are sparse	Strong evidence that representation ranking and tradable predictability can diverge

The broad lesson is consistent across positive and negative results. Transformer components help most when they solve a specific structural problem: changing variable relevance, temporal abstraction, multivariate context, or transfer from broader data. They do not enjoy a universal finance premium.

10. Foundation Models: What Changed After 2025

The arrival of finance-specific foundation models changes the evidence base in a meaningful but limited way. Prior to 2025, the strongest public case for transformers in finance rested heavily on bespoke supervised models and generic time-series architectures. FinCast provides a financial foundation model designed for heterogeneous financial series (Zhu et al., 2025). Kronos goes further by tokenizing candlestick data and pretraining on a multi-market corpus exceeding 12 billion K-line records from 45 exchanges; the AAAI 2026 paper reports gains in price forecasting, volatility forecasting, and synthetic-series fidelity (Shi et al., 2026). These results make it less plausible to argue that financial pretraining is inherently ineffective.

But the economic interpretation remains constrained. A model can forecast levels, volatility, or K-line structure well without producing statistically reliable return forecasts after costs. That is precisely why the 2026 financial-return foundation-model benchmark is important: pretrained models often rank well, but only a small subset of model-asset combinations significantly improve over a random-walk benchmark (Noguer I Alonso and Pereira Franklin, 2026). The correct conclusion is not "foundation models fail." It is that pretraining changes estimation efficiency and representation quality more readily than it changes the underlying information content of returns.

RCIR therefore predicts the largest value from pretraining when one or more of the following conditions hold: target-domain data are scarce; useful structure transfers across assets or domains; the target is representation-rich rather than near-noise raw return; or the model must infer variable/event relevance from many heterogeneous inputs. It predicts less value when the target is close to unpredictable conditional on the available filtration.

11. Institutional Validation and Governance Standard

A credible research architecture should make it hard to accidentally promote a false discovery. Figure 3 shows a minimum evidence ladder.

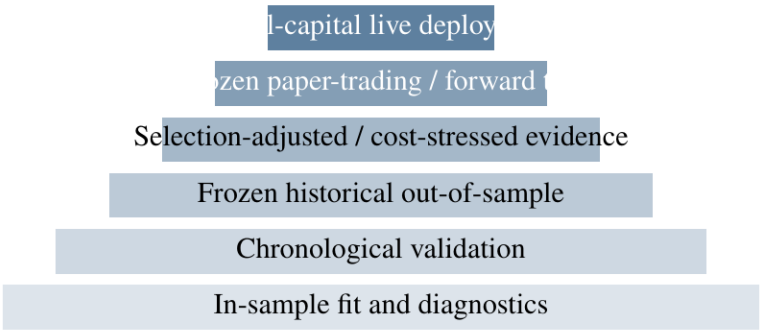

Figure 3: Evidence ladder. Each stage removes a different degree of researcher flexibility. No stage guarantees future performance.

11.1. Minimum promotion gates

Table 6 defines concrete gates. A research program can make these stricter, but making them weaker should require an explicit rationale.

Table 6: Minimum institutional promotion gates

Gate	Question	Minimum evidence
G0: Filtration	Could the result exist only because of data leakage?	Point-in-time reconstruction, vintage/release audit, causal timestamps, survivorship-safe universe
G1: Simplicity	Is the complexity actually necessary?	Linear, tree, recurrent/state-space, and static-attention baselines on identical information
G2: Robustness	Is the result dependent on one seed, window, or cost assumption?	Multiple seeds, subperiods, cost stress, feature-family ablations, parameter perturbations
G3: Selection	Is the result inflated by research search?	Complete trial ledger; Reality Check, DSR, PBO/CSCV, or comparable multiple-testing control
G4: Frozen forward	Does the model survive without redesign?	Frozen code and parameters; paper trading or sequential forward evaluation with prespecified rules
G5: Live scaling	Does implementation preserve the edge?	Small-capital live evidence, slippage attribution, capacity monitoring, drawdown and kill-switch governance

11.2. Research ledger and model cards

Every promoted model should have a research card containing: data lineage, timestamp semantics, intended target, feature families, architecture class, parameter count, training budget, hyperparameter search budget, validation periods, seeds, cost model, baseline set, failed ablations, uncertainty estimates, known failure modes, and conditions that trigger deactivation. This documentation is not bureaucracy; it is part of the experiment.

11.3. Online adaptation deserves special skepticism

A live model that continually updates can react to genuine regime shifts, but it can also chase noise, learn from execution artifacts, or change risk behavior without an auditable promotion event. A safer architecture separates a stable backbone from smaller, tightly governed adaptation components. Forward updates should be versioned and subject to explicit guardrails. Public evidence that uncontrolled continual learning generates persistent live financial alpha remains limited.

12. Interpretability and Failure Analysis

12.1. Interpretability should answer intervention questions

For each purported channel, ask what happens when the channel is altered. If macro-release tokens receive high attention, delay them beyond their actual release time or remove the feature family and observe the forecast change. If cross-asset attention is claimed to capture economic transmission, compare unconstrained attention to a prespecified graph prior and inspect whether the implied edges are stable across seeds. If the model claims long memory, truncate the context and measure state-conditional degradation.

The key question is not "what did the heatmap look like?" It is "does the model's decision change in a stable, economically coherent way when the claimed information source is perturbed?"

12.2. Dominant failure modes

A serious failure taxonomy should include at least:

- revised-data leakage and release-time misalignment;

- label overlap and cross-validation leakage;
- survivorship and constituent-selection bias;
- spurious cross-asset dependence;
- hidden concentration in one crisis or asset;
- seed luck and hyperparameter search inflation;
- underestimated turnover, spread, impact, or financing;
- unstable state estimation;
- attention collapse onto redundant features;
- model complexity that improves fit but not incremental net utility.

These failures should be tested deliberately. A good research process tries to break the model before the market does.

13. Research Architecture Without Revealing a Proprietary Strategy

RCIR can be evaluated without disclosing proprietary securities, features, or execution logic. A defensible non-proprietary architecture is:

1. construct a point-in-time token store with release and vintage metadata;
2. encode economic groups and asset context using only information available at the decision time;
3. estimate a leakage-safe state representation;
4. apply dynamic variable selection before high-capacity sequence modeling;
5. allow temporal routing across multiple horizons;
6. permit cross-asset interaction only through explicit lag and structural constraints;
7. produce probabilistic outputs or calibrated uncertainty, not only point forecasts;
8. pass representations to an independent portfolio layer with hard risk and turnover limits;
9. evaluate the full decision in a simulator with cost and liquidity stress;
10. promote only after frozen forward evidence.

The omitted details—exact securities, feature definitions, vendor mappings, normalization windows, token dictionaries, architecture dimensions, retraining cadence, cost curves, order-generation logic, and risk budgets—are precisely where implementation edge may reside. The scientific claim can still be evaluated through architecture ablations and preregistered tests.

14. Discussion

The upgraded evidence base changes the transformer-in-finance debate in a subtle way. In 2021–2023, the strongest public claim was that attention-based or TFT-style models could be competitive in historical financial trading and

general time-series forecasting. By 2025–2026, finance-specific foundation models demonstrate that large-scale pretraining can learn reusable financial representations, while finance-specific benchmarks simultaneously show that generic transformer leadership does not transfer automatically to risk-adjusted trading performance. The positive and negative results are not contradictory. They point toward a narrower mechanism.

RCIR says the valuable object is not "attention" in isolation. It is adaptive allocation of limited modeling capacity to the information subset that is currently relevant. Variable selection compresses redundant panels. Temporal routing changes the effective memory kernel. Structural masks constrain cross-asset inference. Pretraining supplies a prior when target-domain data are scarce. Probabilistic heads preserve uncertainty. Costs and portfolio constraints keep representation quality from being confused with executable utility.

The formal result makes the same point in minimal form. If the predictive coefficient vector is stable across states, the oracle adaptive model has no population advantage. If the coefficient vector changes, the static model pays a risk penalty equal to the covariance-weighted dispersion in that mapping. The research problem is therefore to determine whether the real market supplies enough state-dependent signal to pay for estimating the more flexible mapping.

This perspective also explains why a transformer can look impressive in one benchmark and weak in another. A model that excels at generic time-series forecasting may be solving a task where long context and nonlinear representation matter. A return-forecasting task may contain almost no predictable information. A systematic-macro task may benefit from cross-asset context but punish unconstrained interaction. The architecture is not portable without the economic problem definition.

15. Limitations and Non-Claims

This paper does not estimate RCIR on a proprietary institutional dataset, and it does not report a new live track record. The propositions are derived in a stylized conditional-linear environment; real financial systems are nonlinear, partially observed, path-dependent, and subject to endogenous execution costs. Proposition 1 establishes the source of population value from state-dependent mapping, not the superiority of a particular neural architecture.

The Regime Reallocation Index is a proposed operationalization, not an established universal regime measure. Different markets may require different target-free shift statistics, and the weights must be frozen ex ante. The literature synthesis includes peer-reviewed papers and recent preprints; evidence from preprints should be treated as provisional until independently replicated. Reported benchmark metrics are study-specific and not directly comparable across datasets, assets, cost assumptions, or research pipelines.

Most importantly, RCIR does not imply that financial returns are highly predictable. It provides a conditional statement: *if* the relevant mapping changes across states and *if* those states can be inferred causally with sufficient sample efficiency, adaptive routing can reduce model misspecification. Either condition can fail.

16. Conclusion

Transformers have a credible but bounded role in institutional quantitative trading. The public evidence through September 21, 2026 does not establish transformers, foundation models, or reinforcement-learning agents as autonomous generators of persistent live alpha. It does show that structured sequence models, dynamic selection, domain-specific pretraining, and constrained cross-asset modeling can improve representation and, in some studies, economic backtest outcomes.

RCIR provides a sharper theoretical explanation for when those improvements should occur. In a stylized setting, the population value of regime-conditioned routing is exactly the covariance-weighted dispersion of the state-specific predictive mapping around the best static mapping. That value is zero when the relation is stable. In finite samples, it must additionally exceed estimation error, research-selection bias, and implementation costs.

The practical research posture follows directly: assume the transformer is unnecessary until it defeats strong alternatives using the same point-in-time information. Require its incremental value to appear in prespecified high-reallocation states, not only in an unconditional average. Constrain cross-asset interaction, preserve release-time semantics, separate representation from portfolio construction, account for realistic costs, record the search process, and reserve the strongest claims for frozen forward and live evidence.

For Ames Investment Systems, the resulting objective is not to build the largest model. It is to build the smallest adaptive information-routing system that can demonstrate incremental net value under an experiment designed to prove it wrong.

A. Proofs

A.1. Proof of Proposition 1

Under Assumption 1,

$$Y_t = \beta_{Z_t}^\top X_t + \varepsilon_t, \qquad \mathbb{E}[\varepsilon_t \mid X_t, Z_t] = 0. \tag{18}$$

The oracle adaptive predictor is $f_A = \beta_{Z_t}^\top X_t$, so its risk is

$$R_A = \mathbb{E}[(Y_t - f_A)^2] = \mathbb{E}[\varepsilon_t^2]. \tag{19}$$

For any static coefficient vector b,

$$\begin{aligned} R_S(b) &= \mathbb{E}[(Y_t - b^\top X_t)^2] && (20)\\ &= \mathbb{E}[((\beta_{Z_t} - b)^\top X_t + \varepsilon_t)^2] && (21)\\ &= \mathbb{E}[((\beta_{Z_t} - b)^\top X_t)^2] + 2\mathbb{E}[\varepsilon_t(\beta_{Z_t} - b)^\top X_t] + \mathbb{E}[\varepsilon_t^2]. && (22) \end{aligned}$$

The cross term is zero by conditional mean independence. Therefore

$$R_S(b) - R_A = \mathbb{E}[(\beta_{Z_t} - b)^\top X_t X_t^\top (\beta_{Z_t} - b)]. \tag{23}$$

Conditioning on Z_t yields

$$R_S(b) - R_A = \mathbb{E}[(\beta_{Z_t} - b)^\top \Sigma_{Z_t} (\beta_{Z_t} - b)]. \tag{24}$$

Choosing $b = \beta^*$, the best static linear predictor, gives Equation (9). Positive semidefiniteness of each Σ_z implies nonnegativity. Equality holds precisely when $\Sigma_{Z_t}^{1/2}(\beta_{Z_t} - \beta^*) = 0$ almost surely, which is equivalent to equality on the covariance-supported subspace. □

A.2. Closed-form static coefficient

When $A = \mathbb{E}[\Sigma_{Z_t}]$ is positive definite, differentiating the static risk gives

$$\beta^* = A^{-1}\mathbb{E}[\Sigma_{Z_t}\beta_{Z_t}]. \tag{25}$$

Thus the best static coefficient is not generally the simple probability-weighted average of regime coefficients. States with greater feature variance in a predictive direction receive more weight.

A.3. Interpretation under equal covariance

If $\Sigma_z = \Sigma$ for all states, then

$$\beta^* = \mathbb{E}[\beta_{Z_t}], \tag{26}$$

and

$$\mathcal{G}_{\text{route}} = \mathbb{E}[(\beta_{Z_t} - \mathbb{E}\beta_{Z_t})^\top \Sigma (\beta_{Z_t} - \mathbb{E}\beta_{Z_t})]. \tag{27}$$

In this special case, routing value is exactly the covariance-weighted variance of the regime-specific predictive coefficients. This is the formal analogue of the paper's qualitative claim that adaptive attention should matter most when information relevance changes across states.

B. Preregistered RCIR Experiment Template

A future empirical paper can preregister the following fields before consuming the frozen test block.

Field	Required declaration
Primary target	Exact forecast or portfolio objective, horizon, and evaluation timestamp
Primary hypothesis	H1–H5 or explicitly defined alternative
Point-in-time contract	Release timestamps, vintage rules, membership handling, missingness, corrections, and corporate actions
Training window	Start/end dates and expanding or rolling rule
Validation window	Dates and tuning budget
Frozen test window	Dates and prohibition on redesign after first consumption
Model ladder	Exact M_0–M_4 implementations and parameter budgets
RRI definition	Components, windows, weights, standardization, and thresholds
Cost model	Spreads, fees, slippage, impact sensitivity, financing, and asset-specific assumptions
Primary statistic	Prespecified forecast loss or net utility measure
Inference	DM/MCS/block bootstrap and trading-specific selection controls
Seed policy	Number of seeds and aggregation rule
Ablations	Dynamic selection, state input, cross-asset mask, pretraining, context length, and cost-aware objective
Promotion rule	Minimum improvement, uncertainty requirement, and forward-test gate
Failure rule	Conditions under which the hypothesis is considered unsupported

C. Non-Proprietary Institutional Research Checklist

A transformer research result should not advance to deployment unless the following questions can be answered affirmatively:

- Is every training and test feature reconstructed point-in-time, including revisions and membership changes?
- Are the same information set and execution timestamps supplied to simpler baselines?

- Is model selection chronological, with all transforms fit only on training data?
- Are overlapping labels purged or embargoed where necessary?
- Are costs applied at the intended decision frequency and stress-tested above the base case?
- Are results stable across seeds, nearby hyperparameters, and multiple subperiods?
- Does no single asset, crisis, or short episode dominate the claimed benefit?
- Do ablations identify incremental value from variable selection, state conditioning, and structural cross-asset constraints?
- Is the experiment count recorded and multiple-testing or backtest-overfitting risk addressed?
- Is the final historical lockbox consumed only once?
- Does a frozen forward test behave within modeled ranges before capital is allocated?
- Is there a documented deactivation rule for distribution shift, cost drift, or abnormal model behavior?

A model that fails this checklist is not necessarily a bad forecasting model. It is simply not yet supported as an institutional trading component.

D. Claim Boundary Matrix

Table 8: What this paper does and does not claim

Claim status	Statement
Formally established in the stylized model	State-conditioned prediction weakly dominates the best static linear predictor in population when the predictive mapping varies across states; the exact risk gap is given by Equation (9).
Supported by current literature	Structured sequence models, variable selection, domain-specific pretraining, and constrained cross-asset mechanisms can improve forecasting or study-specific trading outcomes.
Empirical hypothesis requiring new validation	Incremental transformer value should increase monotonically with a leakage-safe measure of information reallocation such as RRI.
Not established	Transformers or foundation models generate persistent autonomous live alpha across markets.
Not established	Attention weights are causal explanations of market behavior.
Not established	Larger models are better than compact models for institutional trading.

References

Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., et al. (2024). Chronos: Learning the language of time series. *Transactions of Machine Learning Research*. arXiv:2403.07815.

Bailey, D. H., & López de Prado, M. (2014). The deflated Sharpe ratio: Correcting for selection bias, backtest overfitting, and non-normality. *The Journal of Portfolio Management, 40*(5), 94–107. https://doi.org/10.3905/jpm.2014.40.5.094

Bailey, D. H., Borwein, J. M., López de Prado, M., & Zhu, Q. J. (2016). The probability of backtest overfitting. *Journal of Computational Finance, 20*(4), 39–69. https://doi.org/10.21314/JCF.2016.322

Croushore, D., & Stark, T. (2001). A real-time data set for macroeconomists. *Journal of Econometrics, 105*(1), 111–130. https://doi.org/10.1016/S0304-4076(01)00072-0

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics, 13*(3), 253–263. https://doi.org/10.1080/07350015.1995.10524599

Gurkaynak, R. S., Sack, B., & Swanson, E. T. (2005). Do actions speak louder than words? The response of asset prices to monetary policy actions and statements. *International Journal of Central Banking, 1*(1), 55–93.

Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The Model Confidence Set. *Econometrica, 79*(2), 453–497. https://doi.org/10.3982/ECTA5771

Lim, B., Arik, S. O., Loeff, N., & Pfister, T. (2021). Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting, 37*(4), 1748–1764. https://doi.org/10.1016/j.ijforecast.2021.03.012

Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., & Long, M. (2024). iTransformer: Inverted Transformers are effective for time series forecasting. *International Conference on Learning Representations*.

Lo, A. W. (2002). The statistics of Sharpe ratios. *Financial Analysts Journal, 58*(4), 36–52. https://doi.org/10.2469/faj.v58.n4.2453

Nie, Y., Nguyen, N. H., Sinthong, P., & Kalagnanam, J. (2023). A time series is worth 64 words: Long-term forecasting with transformers. *International Conference on Learning Representations*.

Noguer I Alonso, M., & Pereira Franklin, R. (2026). Pretrained time-series foundation models for financial return forecasting. arXiv:2606.27100.

Saly-Kaufmann, A., Wood, K., Peter-Calliess, J., & Zohren, S. (2026). Deep learning for financial time series: A large-scale benchmark of risk-adjusted performance. arXiv:2603.01820.

Shi, X., Wang, S., Nie, Y., Li, D., Ye, Z., Wen, Q., & Jin, M. (2025). Time-MoE: Billion-scale time series foundation models with mixture of experts. *International Conference on Learning Representations*.

Shi, Y., Fu, Z., Chen, S., Zhao, B., Xu, W., Zhang, C., & Li, J. (2026). Kronos: A foundation model for the language of financial markets. *Proceedings of the AAAI Conference on Artificial Intelligence, 40*(30), 25366–25373. https://doi.org/10.1609/aaai.v40i30.39730

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*.

White, H. (2000). A reality check for data snooping. *Econometrica, 68*(5), 1097–1126. https://doi.org/10.1111/1468-0262.00152

Wood, K., Giegerich, S., Roberts, S., & Zohren, S. (2021). Trading with the Momentum Transformer: An intelligent and interpretable architecture. arXiv:2112.08534.

Wood, K., Roberts, S. J., & Zohren, S. (2026). DeePM: Regime-robust deep learning for systematic macro portfolio management. arXiv:2601.05975.

Zeng, A., Chen, M., Zhang, L., & Xu, Q. (2023). Are Transformers effective for time series forecasting? *Proceedings of the AAAI Conference on Artificial Intelligence, 37*(9), 11121–11128. https://doi.org/10.1609/aaai.v37i9.26317

Zhu, Z., Chen, H., Qu, Q., & Chung, V. (2025). FinCast: A foundation model for financial time-series forecasting. *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, 4539–4549. https://doi.org/10.1145/3746252.3761261