History Rhymes: Macro-Contextual Retrieval for Robust Financial Forecasting

Sarthak Khanna[†], Armin Berger[‡‡†§], Muskaan Chopra[†§], David Berghaus[‡‡], Rafet Sifa[‡‡†§]

[‡‡]Fraunhofer IAIS - Department of Media Engineering, Germany
[†]University of Bonn - Department of Computer Science, Germany
[§]Lamarr Institute for Machine Learning and Artificial Intelligence, Germany

***Abstract*—Financial markets are inherently non-stationary: structural breaks and macroeconomic regime shifts often cause forecasting models to fail when deployed out of distribution (OOD). Conventional multimodal approaches that simply fuse numerical indicators and textual sentiment rarely adapt to such shifts. We introduce *macro-contextual retrieval*, a retrieval-augmented forecasting framework that grounds each prediction in historically analogous macroeconomic regimes. The method jointly embeds macro indicators (e.g., CPI, unemployment, yield spread, GDP growth) and financial news sentiment in a shared similarity space, enabling causal retrieval of precedent periods during inference without retraining.**

Trained on seventeen years of S&P 500 data (2007-2023) and evaluated OOD on AAPL (2024) and XOM (2024), the framework consistently narrows the CV to OOD performance gap. Macro-conditioned retrieval achieves the only positive out-of-sample trading outcomes (AAPL: PF = 1.18, Sharpe = 0.95; XOM: PF = 1.16, Sharpe = 0.61), while static numeric, text-only, and naive multimodal baselines collapse under regime shifts. Beyond metric gains, retrieved neighbors form interpretable evidence chains that correspond to recognizable macro contexts, such as inflationary or yield-curve inversion phases, supporting causal interpretability and transparency. By operationalizing the principle that "financial history may not repeat, but it often rhymes," this work demonstrates that macro-aware retrieval yields robust, explainable forecasts under distributional change.

All datasets, models, and source code have been made publicly available.[1]

***Index Terms*—Financial forecasting, Retrieval-Augmented Generation (RAG), Macroeconomic indicators, Regime shifts, Sentiment analysis, Multimodal learning, Out-of-distribution robustness, Time series analysis.**

I. Introduction

Large Language Models (LLMs) are increasingly central to financial analysis, enabling the interpretation of market narratives, extraction of sentiment, and integration of qualitative news with quantitative signals. By distilling insights from vast corpora, LLM-driven systems have advanced forecasting, risk modeling, and economic policy analysis. However, even the most capable models exhibit sharp performance declines when exposed to new market regimes, as distributional changes disrupt learned relationships.

This research has been funded by the Federal Ministry of Education and Research of Germany and the state of North-Rhine Westphalia as part of the Lamarr Institute for Machine Learning and Artificial Intelligence.

[1]https://github.com/sarthak-12/history_rhymes

Financial data are inherently *non-stationary*. Structural breaks, monetary shifts, and macroeconomic shocks continually reshape dependencies between sentiment, volatility, and returns. Models trained on static correlations often perform well in cross-validation (CV) but deteriorate under regime shifts. This exposes a fundamental challenge for LLM-based financial reasoning: *how can we ensure stability and interpretability when markets change?*

We propose that the solution lies in a principle long recognized by economists and historians: *"History doesn't repeat itself, but it often rhymes."* Markets rarely behave identically across time, yet similar macroeconomic conditions tend to produce comparable behavioral and sentiment patterns.

To operationalize this idea, we introduce a **macro-aware retrieval-augmented** framework that grounds each prediction in its most relevant historical precedents. We jointly embed structured macroeconomic indicators (e.g., inflation, unemployment, term spread, GDP growth) and unstructured textual narratives (financial news) into a shared similarity space. During inference, the model retrieves analogous historical contexts and conditions its predictions on them, effectively learning to reason *by precedent* rather than in isolation.

This paradigm offers three key benefits: (i) **adaptive reasoning** through contextual analogy, (ii) **robustness** under out-of-distribution (OOD) conditions, reducing the CV $\rightarrow$ OOD gap in metrics such as F1-score and Sharpe ratio, and (iii) **interpretability** by linking predictions to explicit, human-understandable precedents.

Our work makes three main contributions:

1) We formalize **macro-contextual retrieval** for financial forecasting by unifying macroeconomic indicators and textual embeddings in a single similarity space.
2) We develop a **retrieval-augmented inference architecture** that dynamically incorporates historical analogues to stabilize reasoning across market regimes.
3) We present **empirical results** across multiple assets and macro conditions, showing that contextual retrieval improves both predictive stability and interpretability under distribution shift.

II. Related Work

A. Retrieval-Augmented Generation (RAG) in Finance

The paradigm of **Retrieval-Augmented Generation (RAG)** addresses two key limitations of large language models (LLMs): (i) their reliance on static, closed-book training corpora (leading to outdated or missing information) and (ii) their tendency to hallucinate or fabricate unsupported statements. As defined by AWS, RAG "optimizes the output of a large language model so it references an authoritative knowledge base outside of its training data" before generation. [1] A recent survey by Wu et al. (2024) provides a systematic overview of RAG architectures, including retriever-generator pipelines, reranking, knowledge-base updating and domain-specific adaptations. [2]

In the finance domain, RAG has begun to gain traction. For example, the study by Iaroshev et al. (2024) analyses an RAG system tailored for Q&A over bank quarterly reports and shows that component choice (retriever, embedder, generator) has a large impact on accuracy and relevance. [3] Lee and Roh (2024) study query-expansion, corpus refinement, and long-context management for financial question-answering tasks. [4] The FinDER dataset provides a finance-specific RAG benchmark with expert-annotated query-evidence pairs targeted at real-world financial questions. [5] These contributions demonstrate the growing maturity of RAG for finance: beyond generic retrieval+LLM, research is focusing on domain-tailored embedding, reranking, and dataset engineering.

B. History-Aware and Contextual Retrieval

Standard retrieval in RAG treats each query in isolation, embedding the query and searching a document index based purely on similarity. However, for many real-world tasks, **temporal** or **history-aware** information is crucial. Embedding generation that incorporates prior context (surrounding text or earlier queries) can improve retrieval fidelity. [6] The paper *TimeR^4: Time-aware Retrieval-Augmented Large Language Models* proposes a Retrieve-Rewrite-Retrieve-Rerank pipeline where temporal constraints (via a temporal knowledge graph) are used to improve retrieval. [7] In finance, the temporal dimension is particularly salient (market regimes, macro shifts, seasonality, structural breaks). The benchmark *FinTMMBench: Temporal-Aware Multi-Modal RAG in Finance* explicitly targets retrieval systems that must respect temporal context (news, tables, stock prices over time). [8] These works point to the need for retrieval strategies that go beyond "bag-of-documents": embedding query history, time-stamping vectors, clustering by temporal regimes, or enforcing time-aware reranking. In this sense, our notion of *macro-contextual retrieval*, retrieving historically analogous regimes or macroeconomic contexts rather than just document fragments, aligns with this emerging line of research.

C. Financial Forecasting and Regime-Shift Robustness

Financial forecasting literature has long emphasised the challenge of structural change: parameter instability, regime shifts, concept drift, non-stationarity of economic time-series. Foundational econometric texts (e.g., Hamilton, 1989; Hendry & Mizon, 2001) highlight how forecasting models fail when regime shifts occur after training. [9], [10] More recent surveys show machine learning approaches applied to prediction under regime change: Suárez-Cetrulo et al. (2023) review how ML techniques are being used to tackle structural change and concept drift in finance. [11] Chung et al. (2025) compare GARCH vs deep learning in volatility forecasting under structural breaks and find DL models often outperform when breaks are present. [12] Hybrid modelling approaches (e.g., econometric+ML, ARIMA+LSTM) are also emerging to capture both linear, regime-sensitive structure and nonlinear adaptivity. [13] In parallel, Khanna et al. (2025) propose a unified multimodal financial forecasting framework, which integrates sentiment embeddings and market indicators through cross-modal attention. This work provides evidence that multimodal fusion enhances generalization across market regimes, reinforcing the value of coupling textual and numerical representations in financial prediction [14]. In sum, the forecasting literature emphasises two key gaps relevant to our work: (i) retrieving and conditioning on historical analogues (rather than retraining for every shift) and (ii) aligning text/alternative data with structured numerical/time-series inputs for improved predictive robustness.

D. Interpretability and Grounded Financial LLMs

Interpretability and factual grounding have become central when deploying LLMs in high-stakes domains such as finance. Retrieval acts as one mechanism to anchor LLM outputs to traceable evidence. The domain-specific model *BloombergGPT* demonstrates the value of large-scale pretraining adapted to finance. [15] The benchmark *FinEval* evaluates LLM reasoning and grounding across financial tasks, exposing limitations in vanilla LLMs. [16] Similarly, *FinRAG* applies RAG specifically to financial question answering, showing that retrieval improves factual accuracy and transparency. [17] Our work builds on these interpretability insights, but extends them: by using retrieval not just for textual grounding, but for selecting **macro-contextual, historically analogous regimes** whose structured indicators condition the model's reasoning.

E. Gap and Contribution

While prior work has made important inroads, three key gaps remain, which our approach addresses:

1) **Textual vs. Macro-contextual retrieval:** Most RAG systems in finance target document retrieval for Q&A tasks. They rarely incorporate the retrieval of *historical regimes* or analogues that map past to present.
2) **Temporal/structural conditioning in forecasting:** Forecasting under regime shifts often uses econometric segmentation or ML drift detection, but rarely pairs that with retrieval of contextually analogous periods using retrieval pipelines.
3) **Unified retrieval-generation-forecasting framework:** Prior work mostly stays within Q&A, interpretability, or classification. Our approach unifies retrieval (of macro-contexts + textual evidence) with a generative forecast

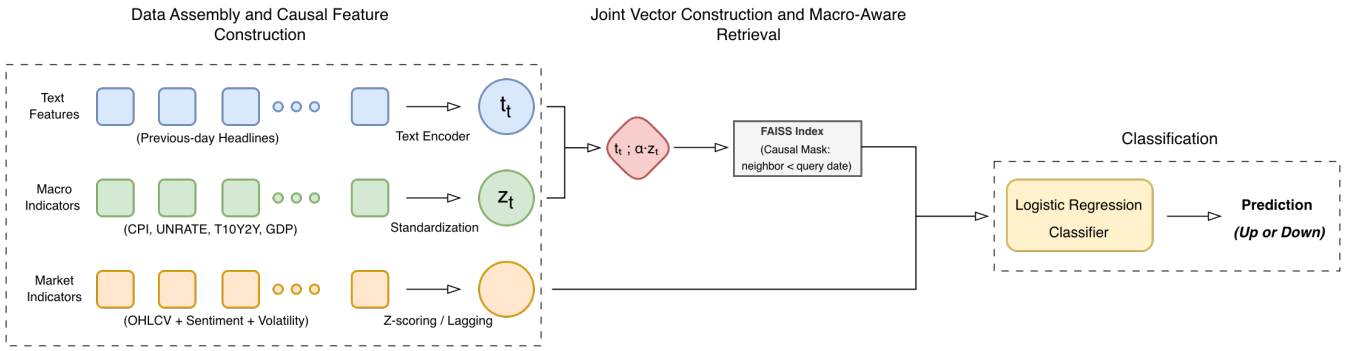


Fig. 1: History-Rhymes pipeline. From left: (top) text embeddings t_t, (middle) macro vectors z_t, (bottom) numeric features $\mathbf{x}_t^{\text{num}}$. The query day forms a fused vector $[t_t; \alpha z_t]$ and searches a FAISS index (built on 2007-2023) under a causal mask (neighbors strictly earlier than the query date). Retrieved neighbors provide a contextual memory r_t which, together with $\mathbf{x}_t^{\text{num}}$, feeds a logistic-regression classifier.

model (LLM conditioned on retrieved context) to deliver adaptive forecasting under distributional shift.

By extending RAG into the *macro-contextual* domain and linking text retrieval + structured-indicator retrieval with forecasting, we provide both an interpretive lens (analogous regime retrieval) and practical adaptivity (no retraining required when regimes evolve).

III. Methodology

A. Data Construction

We build two time-aligned corpora: (i) a training universe of S&P 500 index with trading days from 2007-2023, and (ii) OOD evaluation windows for two stocks, Apple Inc. (AAPL) and Exxon Mobil (XOM) each for the year 2024. For each business day t, we assemble three lagged inputs (to preserve causality):

- **Numerics** $\mathbf{x}_t^{\text{num}}$: OHLCV, lagged returns, realized volatility, and sentiment-derived scalars from previous-day news.
- **Text embedding** $t_t \in \mathbb{R}^d$: a MiniLM encoder fine-tuned on the *Financial PhraseBank* + *FiQA* sentiment dataset from Kaggle[2] (5,842 labeled sentences) to capture domain-specific financial sentiment. Daily headlines for S&P 500 companies are sourced from the *FinSen* dataset[3] (160k articles, 2007-2023). Each previous-day headline is encoded and ℓ_2-normalized to produce t_t.
- **Macro state** $z_t \in \mathbb{R}^p$: CPI, UNRATE, T10Y2Y, and real GDP are derived using the FRED API [4]. Each series is publication-lagged (CPI: 10 bd; UNRATE: 5 bd; GDP: 30 bd), reindexed to business days and forward-filled between releases, then standardized on the 2007-2023 window.

The binary label $y_t \in \{0, 1\}$ indicates whether the next-day open exceeds the previous close.

[2]https://www.kaggle.com/datasets/sbhatti/financial-sentiment-analysis/data?select=data.csv

[3]https://github.com/EagleAdelaide/FinSen_Dataset/tree/main

[4]https://fred.stlouisfed.org/

B. Macro-Contextual Fusion & Query Formation

As described in Fig. 1, we explicitly separate text (t_t) and macro (z_t) channels and construct a fused query used *only* for retrieval:

$$\mathbf{q}_t \;=\; \text{norm}\big([\,t_t\,;\,\alpha\,z_t\,]\big), \tag{1}$$

where α weights macro context (α=0.5 unless noted) and norm$(\cdot)$ denotes ℓ_2 normalization. This aligns narrative tone and macro state in a common similarity space.

C. Causal Retrieval of Historical Analogues

We index all training-period fused vectors $\{\mathbf{q}_\tau\}_{\tau \in \text{2007-2023}}$ with a FAISS inner-product index (cosine after normalization). For an OOD query day t, we enforce a **causal mask** so only $\tau < t$ are eligible. The top-K neighbors $\mathcal{N}_t = \{n_1, \ldots, n_K\}$ are then used to form a *contextual memory* by averaging neighbors' *text* embeddings:

$$r_t \;=\; \frac{1}{K}\sum_{i=1}^{K} t_{n_i}. \tag{2}$$

D. Forecasting Head

The classifier receives the day's numeric features in parallel with the retrieved context:

$$\hat{y}_t \;=\; \sigma\big(W\,[\,\mathbf{x}_t^{\text{num}}\,;\,r_t\,] + b\big), \tag{3}$$

with parameters fit via five-fold `TimeSeriesSplit` on 2007-2023. All scalers, embeddings, and model weights are frozen for OOD tests on AAPL/XOM (2024). This design mirrors the figure: numerics bypass the retriever and provide contemporaneous market microstructure, while r_t supplies regime-aware precedent drawn from macro-conditioned neighbors.

IV. Experiments and Results

A. Experimental Setup

We train on S&P 500 daily data from 2007-2023 and evaluate out-of-distribution (OOD) performance on AAPL (2024) and XOM (2024). The following five configurations are compared, each using identical lag structures, normalization, and causal preprocessing:

- **Numeric-only:** price and volume-based market indicators (OHLCV, volatility, and sentiment-derived numerics) without text or macro inputs.
- **Text-only:** previous-day news embeddings (MiniLM-ft) used independently to capture linguistic sentiment and narrative tone.
- **Multimodal (No-Retrieval):** direct concatenation of numeric, macroeconomic, and textual features into a single multimodal vector.
- **Text-Retrieval (α=0):** retrieval based solely on semantic similarity of news embeddings; macro context ignored in the query.
- **Macro-Retrieval (α=0.5, K=5):** joint text-macro retrieval where similarity search is conditioned on both semantic and macroeconomic state, retrieving the top-K historical analogs under causal masking.

A five-fold `TimeSeriesSplit` preserves chronological order, ensuring each validation segment only accesses past data. During both cross-validation (CV) and OOD evaluation, retrieval operates strictly on training periods: for each validation or test day, its joint query vector $\mathbf{q}_t = [t_t; \alpha z_t]$ searches a FAISS index containing *only* earlier dates, enforcing causal retrieval. For OOD tests, all scalers, encoders, the FAISS index, and classifier weights learned on 2007-2023 are frozen. We build a dense similarity index using **FAISS** (Facebook AI Similarity Search), a high-performance library for nearest-neighbor retrieval that enables efficient similarity search over millions of high-dimensional embedding vectors on CPU or GPU.

Representation diagnostics (t-SNE): To qualitatively assess how well the learned joint representations capture regime structure, we employ t-SNE to project the high-dimensional (Text + Macro) embeddings into two dimensions. These projections visualize how training and test samples distribute across macroeconomic conditions and help confirm that retrieved neighbors lie in historically plausible regions of the embedding space, reflecting semantic and economic similarity.

B. Evaluation Metrics

We report both *classification* and *financial* metrics, along with two robustness deltas.

- **Classification:** Accuracy, Precision, Recall, F_1, AUROC, and MCC.
- **Financial:** Win Rate, Profit Factor (PF), and annualized Sharpe ratio (Sharpe_{252}) under long-short daily positions $p_t \in \{-1, +1\}$.
- **OOD robustness deltas:**

$$\Delta F_1 = F_1^{\text{CV}} - F_1^{\text{OOD}}, \quad \Delta\text{Sharpe} = \text{Sharpe}^{\text{CV}} - \text{Sharpe}^{\text{OOD}}$$

C. Cross-Validation (2007-2023)

Tables I and II summarize in-sample (CV) performance under a long-short trading setup. Here, retrieval occurs within each fold's training window, each validation day retrieves analogs only from earlier periods, maintaining strict temporal causality. Thus, CV evaluates how well the model generalizes to *future but still in-sample regimes*, such as post 2008 recovery or the COVID-19 period, while the OOD tests (AAPL and XOM 2024) later measure full regime transfer.

Across metrics, **Numeric-only** achieves the highest AUROC (0.82) and Sharpe (1.99), confirming that market microstructure features remain strong in stationary conditions. **Multimodal (No-Ret)** yields the best F_1 (0.74), showing that fusing text and macro data enhances next-day classification accuracy. Both retrieval variants slightly reduce peak F_1 but yield smoother financial performance: **Macro-Retrieval** (α=0.5) attains Win Rate 0.52, PF 1.40, Sharpe 1.55, and MCC 0.23, implying a more stable signal-to-noise balance.

This pattern reflects the role of retrieval in CV: rather than overfitting to short-term correlations, the retriever draws from historically analogous days within the training years, often from similar macroeconomic states, acting as a temporal regularizer. Even when evaluated on unseen segments within 2007-2023, macro-aware retrieval preserves profitability and signal coherence, suggesting robustness to mild within sample regime shifts. Text-only and numeric baselines remain more volatile: the former underperforms due to language noise, while the latter overfits to transient technical patterns. By conditioning similarity on macro context, Macro-Retrieval mitigates both issues, balancing accuracy and financial consistency across folds.

TABLE I: Cross-Validation (2007-2023): Classification Metrics (long-short). Best results are in **bold**.

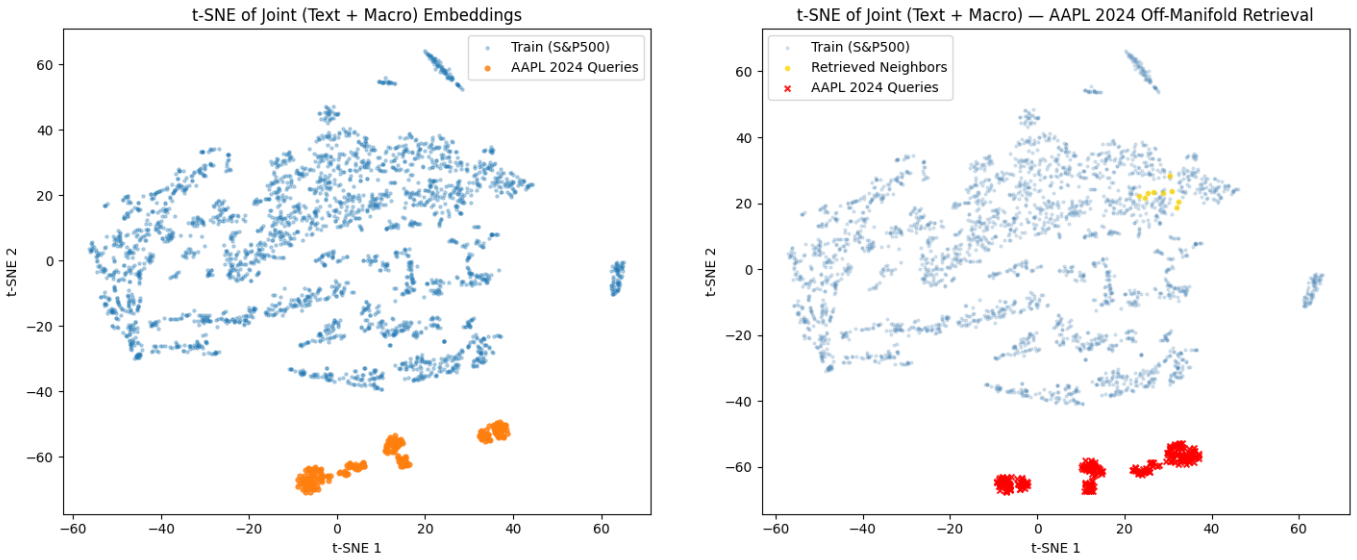
Setting	Acc	F1	MCC	AUROC	Prec	Rec
Numeric-only	0.62	0.71	**0.28**	**0.82**	**0.65**	0.88
Text-only	0.55	0.67	0.01	0.50	0.56	0.87
Multimodal (No-Ret)	**0.63**	**0.74**	0.27	0.73	0.61	**0.96**
Text-Retrieval (α=0)	0.60	0.70	0.22	0.77	0.62	0.88
Macro-Retrieval (α=0.5)	0.61	0.73	0.23	0.72	0.60	**0.96**

TABLE II: Cross-Validation (2007-2023): Financial Metrics (long-short). Best results are in **bold**.

Setting	Profit Factor	Win Rate	Sharpe (252)
Numeric-only	**1.65**	**0.54**	**1.99**
Text-only	0.94	0.48	-0.37
Multimodal (No-Ret)	1.48	0.53	1.91
Text-Retrieval (α=0)	1.50	0.53	1.55
Macro-Retrieval (α=0.5)	1.40	0.52	1.55

D. Out-of-Distribution: AAPL (2024)

AAPL 2024 queries occupy a distinct off-manifold region relative to the 2007-2023 training distribution (Fig. 2a). The t-SNE projection of joint (Text + Macro) embeddings shows that AAPL 2024 points (orange clusters) are well separated from the S&P 500 training cloud (blue), indicating a substantial regime and semantic shift. Despite this displacement, retrieval-based methods successfully locate relevant historical analogs in coherent regions of the training manifold (yellow points in Fig. 2b), validating that the FAISS index retrieves economically



(a) t-SNE (Text+Macro): Train (blue) vs. AAPL 2024 queries (orange).

(b) AAPL 2024 off-manifold retrieval: retrieved neighbors (yellow) from comparable macro regimes.

Fig. 2: AAPL representation diagnostics and retrieval results.

and linguistically plausible precedents rather than random matches.

Static baselines - Numeric-only, Text-only, and Multimodal (No-Retrieval) all degrade sharply under this OOD regime (Tables III, IV). Their Profit Factors collapse toward unity (PF $\approx$ 1.0), and Sharpe ratios turn negative, indicating risk-adjusted losses. Numeric-only retains modest classification accuracy (F_1 0.62, AUROC 0.57) due to strong price momentum, but this signal fails to convert into profitable trades. Text-only and Multimodal (No-Ret) struggle to generalize: textual sentiment patterns from 2024 news diverge from the linguistic styles seen in 2007-2023, while concatenated fusion lacks the ability to adapt to these shifts.

Retrieval variants show improved resilience. Text-Retrieval (α=0) modestly lifts AUROC (0.52) but remains unprofitable (Sharpe -1.63), confirming that text-based analogy alone cannot capture changing macro context. In contrast, Macro-Retrieval (α=0.5, K=5) maintains stable classification and the only positive financial outcome: Profit Factor 1.18, Win Rate 0.47, Sharpe 0.95. This suggests that constraining similarity search by macroeconomic conditions enables retrieval of historically comparable periods, such as inflationary or yield curve inversion phases, thus restoring interpretability and profitability even when the query year lies outside the training manifold.

The robustness plots (Fig. 3a, 3b) corroborate this behavior. Macro-Retrieval exhibits the smallest F_1 and Sharpe degradation between CV and OOD ($\Delta F_1 \approx 0.24$, ΔSharpe ≈ 0.60), while all other configurations suffer substantial drops (Numeric-only ΔSharpe $\approx$3.6). This narrow generalization gap demonstrates that macro-conditioned retrieval mitigates regime sensitivity and preserves both signal coherence and trading stability when facing unseen macro-financial conditions.

TABLE III: OOD (AAPL 2024): Classification Metrics. Best results are in **bold**; where a metric saturates (e.g., Recall = 1), the next highest meaningful value is highlighted.

Setting	Acc	F1	MCC	AUROC	Prec	Rec
Numeric-only	**0.48**	**0.62**	**0.15**	**0.57**	**0.45**	**0.98**
Text-only	0.43	0.61	0.00	0.43	0.43	1.00
Multimodal (No-Ret)	0.47	0.55	0.01	0.50	0.44	0.74
Text-Retrieval (α=0)	0.43	0.59	-0.05	0.52	0.43	0.94
Macro-Retrieval (α=0.5)	0.45	0.49	-0.06	0.50	0.41	0.61

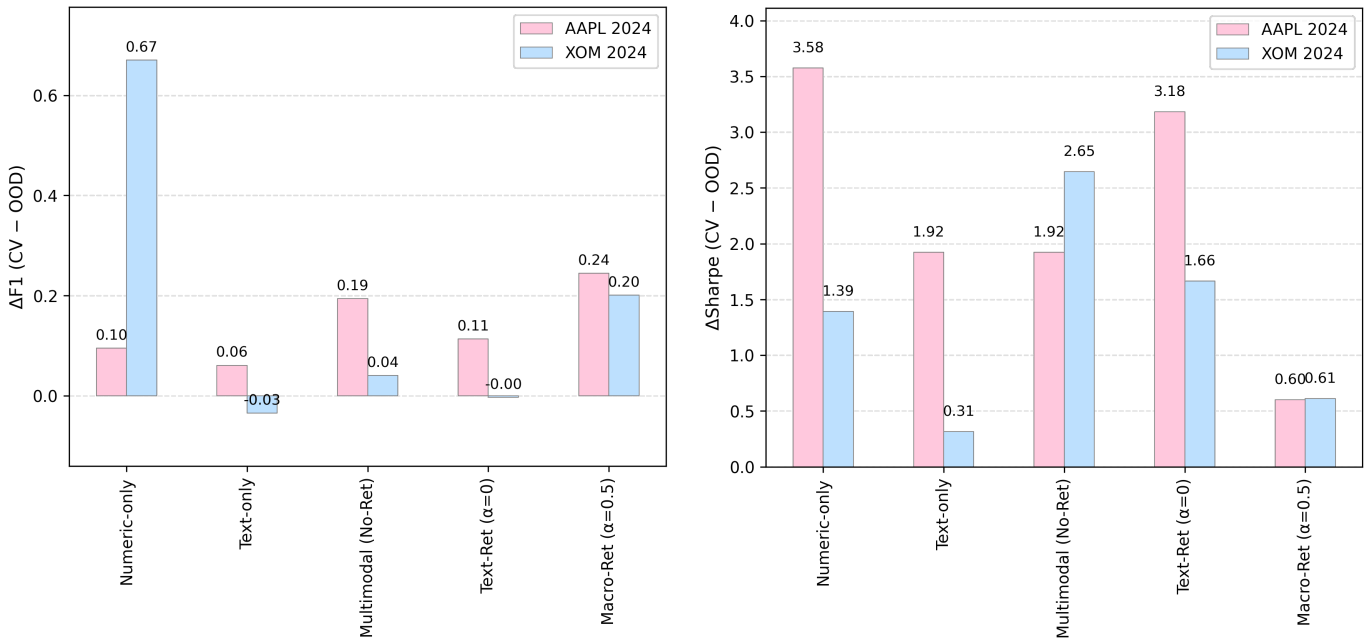
TABLE IV: OOD (AAPL 2024): Financial Metrics. Best results are in **bold**.

Setting	Profit Factor	Win Rate	Sharpe (252)
Numeric-only	0.76	0.41	-1.58
Text-only	0.67	0.39	-2.30
Multimodal (No-Ret)	1.00	0.44	-0.01
Text-Retrieval (α=0)	0.75	0.40	-1.63
Macro-Retrieval (α=0.5)	**1.18**	**0.47**	**0.95**

E. Cross-Asset Generalization: XOM (2024)

Using the same frozen pipeline trained on 2007-2023 data, XOM 2024 exhibits a similar off-manifold structure in the joint (Text + Macro) embedding space (Fig. 4a). The t-SNE projection shows that XOM 2024 queries (red) cluster in a compact region clearly separated from the S&P 500 training manifold (blue), indicating a structural distribution shift in both narrative style and macroeconomic context. Nevertheless, Macro-Retrieval successfully identifies coherent historical analogs (yellow clusters in Fig. 4b), confirming that causal retrieval recovers training periods with comparable macro regimes rather than random or semantically unrelated matches.

Under this cross-asset shift, static baselines lose generalization strength. Numeric-only, despite yielding the weakest classification signal (F_1 0.04, AUROC 0.58), happens to



(a) F_1 drop for AAPL and XOM.

(b) Sharpe ratio drop for AAPL and XOM.

Fig. 3: Robustness degradation across CV and OOD. Macro-Retrieval exhibits the smallest performance drop.

achieve a moderate Profit Factor (1.10) and Sharpe (1.39), suggesting that short-term price momentum patterns still partially align between training and 2024 energy-sector data. In contrast, Text-only and Multimodal (No-Ret) configurations attain superficially high F_1 scores (0.70) due to overconfident recall but fail to translate this into profitable trading (PF < 1), reflecting semantic drift in 2024 financial language. Their high variance in returns indicates a mismatch between textual sentiment shifts and realized market movements.

Retrieval-enhanced approaches again demonstrate stronger cross-asset robustness. Text-Retrieval (α=0) slightly improves classification (F_1 0.71) but remains financially unstable (Sharpe 1.66, PF 0.98), whereas Macro-Retrieval (α=0.5) maintains balanced performance across both classification and financial dimensions. It achieves PF 1.16, WinRate 0.52, and a positive Sharpe 0.61 (Table VI), marking the only configuration with consistent profitability under causal, frozen OOD testing. This suggests that the macro-conditioned retriever generalizes not only across time but also across *assets*, retrieving S&P 500 days characterized by similar macroeconomic conditions, such as commodity price shocks or yield curve tightening, that align with XOM's 2024 environment.

The robustness plots (Fig. 3a, 3b) reinforce this conclusion. Macro-Retrieval yields the smallest degradation in both F_1 and Sharpe between CV and OOD ($\Delta F_1 \approx 0.20$, ΔSharpe ≈ 0.61), while all other configurations experience much larger volatility-induced collapses. These findings support the hypothesis that macro-aware retrieval provides genuine cross-asset transferability, retaining profitability when textual and sector, specific patterns diverge but macroeconomic drivers remain comparable.

TABLE V: OOD (XOM 2024): Classification Metrics. Best results are in **bold**.

Setting	Acc	F1	MCC	AUROC	Prec	Rec
Numeric-only	0.46	0.04	0.02	0.58	**0.60**	0.02
Text-only	0.54	0.70	0.00	0.53	0.54	1.00
Multimodal (No-Ret)	0.54	0.70	0.07	**0.60**	0.54	1.00
Text-Retrieval (α=0)	**0.55**	**0.71**	**0.10**	0.55	0.55	**0.99**
Macro-Retrieval (α=0.5)	0.51	0.53	0.03	0.54	0.55	0.52

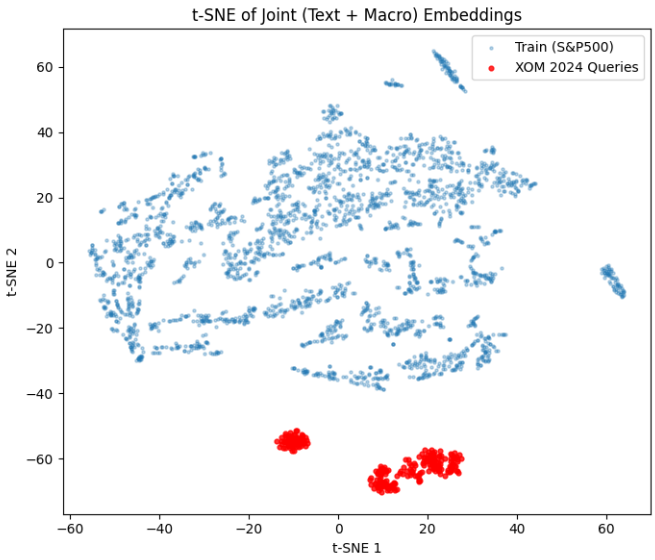
TABLE VI: OOD (XOM 2024): Financial Metrics (long-short). Best results are in **bold**.

Setting	Profit Factor	Win Rate	Sharpe (252)
Numeric-only	1.10	0.54	1.39
Text-only	0.89	0.45	0.31
Multimodal (No-Ret)	0.89	0.45	**2.65**
Text-Retrieval (α=0)	0.98	0.46	1.66
Macro-Retrieval (α=0.5)	**1.16**	**0.52**	0.61

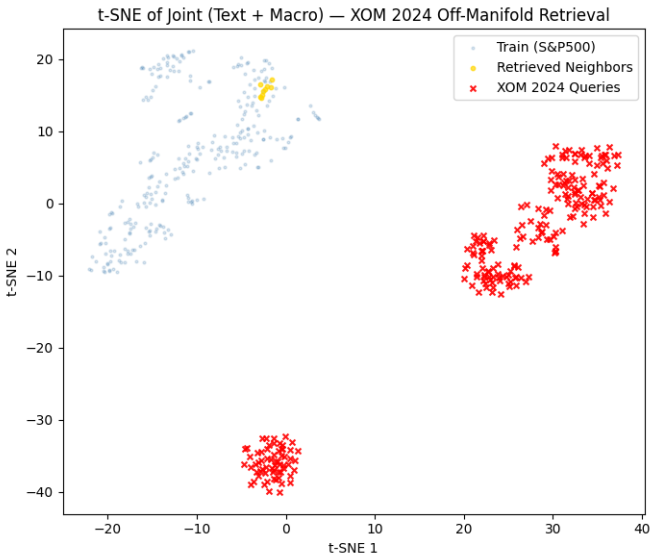
F. Summary

Across both AAPL and XOM 2024 evaluations, the joint Text + Macro embedding space confirms a clear distributional shift: OOD queries occupy compact, low-variance regions well separated from the 2007-2023 training manifold. Despite this shift, macro-conditioned retrieval consistently delivers the most stable generalization, exhibiting the smallest robustness degradations in ΔF_1 and ΔSharpe across assets. It is also the *only* configuration achieving simultaneous profitability and positive risk-adjusted returns (PF > 1, Sharpe > 0) under long-short trading on AAPL 2024, and maintaining stability under cross-asset transfer to XOM 2024.

The t-SNE diagnostics provide complementary qualitative evidence: retrieved neighbors occupy coherent, economically meaningful regions of the training distribution-often aligned



(a) t-SNE (Text+Macro): Train (blue) vs. XOM 2024 queries (red).



(b) XOM 2024 off-manifold retrieval: retrieved neighbors (yellow) from comparable macro regimes.

Fig. 4: XOM representation diagnostics and retrieval results.

with inflationary, energy-shock, or inverted-yield regimes-demonstrating that the model retrieves analogs from historically comparable macro conditions rather than relying on superficial textual overlap. Together, these findings confirm that macro-aware retrieval not only mitigates regime sensitivity but also offers interpretable, causally grounded evidence for the "history rhymes" hypothesis in financial forecasting.

V. Discussion

Our findings demonstrate that *macro-contextual retrieval* consistently enhances robustness under distribution shift, a long-standing challenge in econometrics and financial machine learning. Unlike conventional multimodal fusion approaches [16], [18] that concatenate features without temporal or macroeconomic grounding, our method anchors each forecast in historically analogous regimes. This design aligns with emerging history-aware retrieval paradigms in language modeling [6], extending them into the econometric domain through causal, macro-conditioned search.

TABLE VII: AAPL (2024) → Historical Analogues: Similarity and Macro Distance Metrics.

Query → Neighbor	sim_joint	sim_text	macro_L2	Neighbor ID
2024-01-18 → 2020-01-31	0.9657	0.9657	0.4850	1
2024-01-18 → 2018-07-19	0.9649	0.9649	0.5306	2
2024-01-18 → 2018-08-14	0.9663	0.9663	0.5365	3
2024-02-20 → 2020-01-31	0.9654	0.9657	0.6522	1
2024-02-20 → 2018-07-19	0.9644	0.9648	0.6943	2
2024-02-20 → 2018-08-14	0.9658	0.9663	0.7140	3
2024-09-09 → 2018-07-19	0.9652	0.9650	0.2766	1
2024-09-09 → 2018-09-04	0.9655	0.9654	0.3025	2
2024-09-09 → 2020-01-31	0.9660	0.9659	0.3220	3

TABLE VIII: Illustrative "History Rhymes" examples: retrieved historical analogues where textual semantics and macro regimes coincide for **AAPL (2024)** queries.

2024 Query Headline	**Retrieved Historical Headline**	**Year**
iPhone 16 To Feature Arm's Next-Gen AI Chip Technology, What You Should Know About Apple's Move	Dollar Rises on Trade War Concerns	2018
Apple Inc. (AAPL) Is Attracting Investor Attention: Here Is What You Should Know	US 10Y Bond Yield Hits 16-Week Low	2020
Seizing India's Investment Opportunities In 2024 With Octa	US Small Business Optimism Nears Survey High in July	2018

What the diagnostics reveal.: t-SNE confirms that 2024 queries lie off the 2007-2023 manifold, while drop plots show smaller CV→OOD degradations for Macro-Retrieval.[5] More importantly, Tables VII-VIII make this mechanism explicit. Each query-neighbor pair is evaluated using three metrics: (i) **sim_text**, the cosine similarity between their news embeddings; (ii) **sim_joint**, the similarity of the fused vectors $[t_t; \alpha z_t]$ used for retrieval; and (iii) **macro_L2**, the Euclidean distance between their standardized macro feature vectors (z_t). High sim_joint values, jointly influenced by textual and macro similarity, and low macro_L2 distances indicate that retrieved neighbors not only share linguistic context but also reside in comparable macroeconomic regimes (e.g., 2018 trade-war period, early 2020 rate cuts). This dual constraint helps convert off-manifold queries into *economically meaningful analogies*-retrieved days that resemble the query both in narrative tone and in macroeconomic context, rather than superficial text matches.

[5]See Figs. 2, 4, 3a, and 3b.

Why trading metrics matter.: As shown in Table VII, high sim_text and sim_joint values capture strong semantic alignment, while low macro_L2 distances ensure that retrieved neighbors occur under comparable macroeconomic conditions. Together with the headline correspondences in Table VIII, this dual filtering prevents purely textual but economically irrelevant matches. Consequently, Macro-Retrieval achieves PF$>$1 and positive Sharpe on AAPL 2024 while remaining competitive on AUROC and MCC, demonstrating that profitability arises when semantic similarity is reinforced by macro consistency, not by language overlap alone.

Positioning vs. prior work.: Table IX situates our framework among multimodal and retrieval-augmented forecasting systems. Unlike prior RAG applications that emphasize Q&A or document grounding and rarely report daily OOD trading metrics, our macro-conditioned retrieval explicitly targets regime-aware forecasting and is evaluated with portfolio-relevant criteria (PF, Sharpe) under frozen out-of-sample tests.

TABLE IX: Comparison with prior retrieval and multimodal forecasting studies.

Aspect	Ours CV (2007-2023)	Ours OOD AAPL 2024	FinSeer / Finrag [17], [19]
Modalities	Text + Numerics + Macro (kNN)	Text + Numerics + Macro (kNN)	Numeric + LLM signals
Retrieval	Yes (macro-cond.)	Yes (macro-cond.)	Yes (domain retriever)
Horizon	Daily S&P 500 train history	Daily single stock OOD	Daily; multi-dataset
Reported Metrics	F1, PF, Sharpe, AUROC	F1, PF, Sharpe, AUROC	Accuracy

A. Limitations and Future Work

Static retriever. The FAISS index remains fixed post-training; adaptive or reinforcement-tuned retrieval could dynamically track evolving macro conditions.

Scope. We focus on U.S. equities; extending to multi-country universes, policy-sensitive assets, or alternative macro calendars would help test cross-regime generality further.

Strategy realism. Our evaluation assumes frictionless daily rebalancing; incorporating transaction costs, slippage, and portfolio constraints would enable operational deployment.

VI. Conclusion

We presented a *macro-contextual retrieval* framework that anchors each forecast in historically analogous economic regimes by jointly encoding daily news and macro indicators. Across seventeen years of S&P 500 training data and out-of-distribution evaluations on AAPL 2024 and XOM 2024, the proposed macro-conditioned retrieval consistently preserved profitability and risk-adjusted performance, whereas static numeric, text-only, and naive multimodal baselines deteriorated under regime shift. Beyond metric improvements, the retrieved neighbors form interpretable evidence chains that correspond to recognizable macro episodes, such as high inflation phases and yield-curve inversions, thereby enhancing transparency and causal interpretability. Financial history may not repeat, but it often *rhymes*; harnessing those rhymes through macro-aware retrieval offers a principled path towards more robust and explainable daily forecasting under distributional change.

References

[1] Amazon Web Services, "What is retrieval-augmented generation (rag)?" https://aws.amazon.com/what-is/retrieval-augmented-generation/, 2024, accessed 2025-10-24.

[2] S. Wu, Y. Xiong, Y. Cui, H. Wu, C. Chen, Y. Yuan, L. Huang, X. Liu, T.-W. Kuo, N. Guan *et al.*, "Retrieval-augmented generation for natural language processing: A survey," *arXiv preprint arXiv:2407.13193*, 2024.

[3] I. Iaroshev, R. Pillai, L. Vaglietti, and T. Hanne, "Evaluating retrieval-augmented generation models for financial report question and answering." *Applied Sciences (2076-3417)*, vol. 14, no. 20, 2024.

[4] J. Lee and M. Roh, "Multi-reranker: Maximizing performance of retrieval-augmented generation in the financerag challenge," *arXiv preprint arXiv:2411.16732*, 2024.

[5] C. Choi, J. Kwon, J. Ha, H. Choi, C. Kim, Y. Lee, J.-y. Sohn, and A. Lopez-Lira, "Finder: Financial dataset for question answering and evaluating retrieval-augmented generation," *arXiv preprint arXiv:2504.15800*, 2025.

[6] Anthropic, "Contextual retrieval: Improving retrieval quality by using query history and surrounding context," https://www.anthropic.com/..., 2024, accessed 2025-10-24.

[7] X. Qian, Y. Zhang, Y. Zhao, B. Zhou, X. Sui, L. Zhang, and K. Song, "Timer4: Time-aware retrieval-augmented large language models for temporal knowledge graph question answering," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 6942–6952.

[8] F. Zhu, J. Li, L. Pan, W. Wang, F. Feng, C. Wang, H. Luan, and T.-S. Chua, "Fintmmbench: Benchmarking temporal-aware multi-modal rag in finance," *arXiv preprint arXiv:2503.05185*, 2025.

[9] J. D. Hamilton, "A new approach to the economic analysis of nonstationary time series and the business cycle," *Econometrica: Journal of the econometric society*, pp. 357–384, 1989.

[10] D. F. Hendry and G. E. Mizon, "Forecasting in the presence of structural breaks and policy regime shifts," *Identification and inference for econometric models*, pp. 480–502, 2005.

[11] A. L. Suárez-Cetrulo, D. Quintana, and A. Cervantes, "Machine learning for financial prediction under regime change using technical analysis: A systematic review," 2023.

[12] V. Chung, J. Espinoza, and R. Quispe, "Forecasting financial volatility under structural breaks: A comparative study of garch models and deep learning techniques," *Journal of Risk and Financial Management*, vol. 18, no. 9, p. 494, 2025.

[13] D. Stempień and R. Ślepaczuk, "Hybrid models for financial forecasting: Combining econometric, machine learning, and deep learning models," *arXiv preprint arXiv:2505.19617*, 2025.

[14] S. Khanna, A. Berger, D. Berghaus, T. Deusser, L. Sparrenberg, and R. Sifa, "Towards unified multimodal financial forecasting: Integrating sentiment embeddings and market indicators via cross-modal attention," *arXiv preprint arXiv:2508.13327*, 2025.

[15] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, and G. Mann, "Bloomberggpt: A large language model for finance," *arXiv preprint arXiv:2303.17564*, 2023.

[16] L. Zhang, W. Cai, Z. Liu, Z. Yang, W. Dai, Y. Liao, Q. Qin, Y. Li, X. Liu, Z. Liu *et al.*, "Fineval: A chinese financial domain knowledge evaluation benchmark for large language models," *arXiv preprint arXiv:2308.09975*, 2023.

[17] K. Kannammal, M. K, K. P, G. G, and A. C, "Fin-rag a rag system for financial documents," *International Journal of Innovative Science and Research Technology*, pp. 1761–1767, 04 2025.

[18] M. Xiao, Z. Jiang, L. Qian, Z. Chen, Y. He, Y. Xu, Y. Jiang, D. Li, R.-L. Weng, M. Peng *et al.*, "Enhancing financial time-series forecasting with retrieval-augmented large language models," *arXiv preprint arXiv:2503.67890*, 2025.

[19] M. Xiao, Z. Jiang, L. Qian, Z. Chen, Y. He, Y. Xu, Y. Jiang, D. Li, R.-L. Weng, M. Peng, and coauthors, "Retrieval-augmented large language models for financial time series forecasting," *arXiv preprint arXiv:2502.05878*, 2025.