Leakage-Aware Benchmarking of LLM Forecasting: Real-Time Nowcasts as the Decision-Time Input for Macro Factor Ranking

Mao Guan [1] **Qian Chen** [1]

Abstract

Forecasting benchmarks for retrieval-augmented LLMs routinely confound model capability with information leakage: features labeled with a target's timestamp are often not observable at the system's decision time. We study leakage-controlled equity factor ranking with a retrieval-augmented 7B open-source LLM forecaster. At each month-end from 2023-04 to 2026-03, the forecaster observes only decision-time information: lag-shifted FRED macro variables, recent macro-event summaries, and the Cleveland Fed's archived daily CPI nowcast for unreleased current-month inflation. A macro-analog retrieval module selects historical states, a critic LLM compresses them into one tactical rule, and an actor LLM maps the current state and recent rules into scores for seven U.S. equity style factors. The full pipeline obtains a median monthly Spearman rank IC of $+0.154$, with positive means across three non-overlapping contiguous 12-month subwindows; the mean IC remains statistically underpowered, with a bootstrap 95% confidence interval that includes zero. Non-LLM baselines under the same decision-time constraint demonstrate that a kNN macro-analog model recovers a comparable median IC, indicating that real-time inflation information and macro-similar retrieval explain much of the median signal. The LLM pipeline retains higher mean IC and a stronger long-short allocation sanity check, suggesting that any marginal benefit is concentrated in the extreme rankings that drive long-short portfolio formation. A descriptive audit of the 36 critic rules and per-month case studies appears in the appendix.

[1]Independent Researcher. Correspondence to: Mao Guan <guanmao771@gmail.com>.

Accepted at the ICML 2026 Workshop on AI Forecasting (non-archival).

1. Introduction

Forecasting systems are often evaluated under a fiction of real-time observability. In macro-driven financial forecasting, this fiction is consequential: the CPI value labeled as month t is typically not available at the month-end decision date for month t: the U.S. Bureau of Labor Statistics (BLS) publishes it about ten days later. A walk-forward backtest that reads this value directly leaks future information (Bailey & Lopez de Prado, 2014).

This issue is amplified in LLM and retrieval-augmented forecasting pipelines (Lopez-Lira & Tang, 2023; Xiao et al., 2024), and recent benchmarks have begun to evaluate LLM-based agents on financial decision-making tasks (Li et al., 2025). Such systems ingest macro narratives, event summaries, and historical analogs, but these inputs are only meaningful if they respect what a decision-maker could have observed at the time. We therefore ask whether a 7B retrieval-augmented LLM forecaster can rank equity style factors under a strict decision-time information constraint, and how much of any observed signal is attributable to the LLM architecture versus to the underlying decision-time information set (Chen et al., 2025).

We build a leakage-controlled monthly benchmark for ranking seven U.S. equity style factors. The key design choice is to replace unreleased current-month CPI with the Cleveland Fed's archived daily inflation nowcast (Knotek & Zaman, 2017; Federal Reserve Bank of Cleveland, 2026), while shifting FRED CPI and unemployment by one month. A macro-analog retrieval module selects historical states with a strict ≥ 12-month embargo, a critic LLM writes one tactical rule from the analogs (Shinn et al., 2023), and an actor LLM maps the current real-time state into factor scores. The 2023-04 to 2026-03 evaluation window is set by the joint availability of factor returns, lag-corrected FRED macro inputs, the Cleveland Fed nowcast archive, and the event corpus.

Our contributions are fourfold. *(i)* We construct a leakage-controlled walk-forward benchmark for macro-conditioned equity factor ranking that respects publication lags throughout (Harvey et al., 2016). *(ii)* We introduce a macro-RAG actor-critic forecaster that compresses retrieved historical

analogs into tactical rules before scoring factors, runnable with 4-bit quantization on commodity GPU hardware. *(iii)* Under the same decision-time information set, we disentangle the contributions of model class and decision-time information: a kNN macro-analog baseline recovers a comparable median rank IC to the full LLM pipeline, while the LLM's residual edge concentrates in mean IC and a long-short allocation sanity check that are driven by extreme-ranking months and do not reach conventional significance at $n = 36$. *(iv)* We release the 36 reproducible critic rules, the parsed Cleveland Fed nowcast archive frozen on 2026-05-13, and representative case studies, alongside the walk-forward harness.

2. Setup

Factors. Seven cross-sectional U.S. equity style factors: SMB, HML, RMW, CMA (Fama & French, 2015), UMD (Jegadeesh & Titman, 1993), BAB (Frazzini & Pedersen, 2014), and QMJ (Asness et al., 2019). Sources: the Kenneth French Data Library (French, 2026) and the AQR Capital Management public factor library (AQR Capital Management, 2026). The market factor (MKT_RF) is excluded from prediction.

Label. The label at month t is the cross-sectional rank of the market-adjusted residual return $r^{\text{resid}}_{f,t} := r_{f,t} - \beta_{f,t}\, r_{\text{MKT},t}$, where $\beta_{f,t}$ is the rolling 60-month OLS beta estimated strictly with data $s < t$.

Decision-time inputs. All methods use decision-time-clean inputs: FRED CPI YoY, CPI level, and Unemployment shifted by one calendar month, plus the daily Cleveland Fed CPI YoY nowcast queried at date t. All remaining market variables from FRED, such as term spread and VIX, are treated as observable at t without publication lag. The non-LLM baselines use only this structured macro / nowcast subset; the LLM actor additionally receives decision-time event summaries from recent FOMC and CPI releases.

Walk-forward. 36 monthly decisions, 2023-04-30 through 2026-03-31.

Metric. Per-month Spearman rank IC between the model's seven scores and the realised seven ranks of $r^{\text{resid}}_{f,t}$. We report mean, median (outlier-robust), three contiguous 12-month sub-window means, bootstrap 95% CI on the mean (10,000 resamples, seed 42), and a sign-symmetric permutation p-value (10,000 permutations). With $n = 36$ and per-month IC standard deviation ≈ 0.45 (close to the Spearman null when ranking seven items), single-window tests have low statistical power at the 5% level for any modest signal.

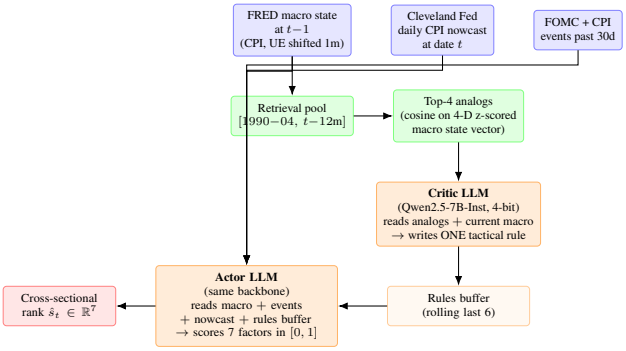

Figure 1. Leakage-controlled monthly decision pipeline. At each month-end t, the system constructs only information observable at t: lag-shifted FRED variables, archived Cleveland Fed CPI nowcasts, and recent macro-event summaries. Historical analogs are retrieved from months at least 12 months before t. A critic compresses the analog evidence into one rule; the actor maps the current state and recent rules into seven factor scores.

3. Method

Architecture. The pipeline is a two-LLM single-backbone actor-critic with the critic conditioned on retrieved macro-analogous history (Laskin et al., 2023) (Figure 1). Both LLMs are Qwen2.5-7B-Instruct (Qwen Team, 2024) in 4-bit NF4 quantization via bitsandbytes (Dettmers et al., 2023), greedy decoding ($T = 0$); month-by-month outputs are deterministic.

Monthly decision protocol. For each month-end t we *(i)* construct a real-time macro state using lag-shifted FRED variables and the archived CPI nowcast at t; *(ii)* retrieve top-K analog months ($K = 4$, matching the four quadrants of the investment clock, Greetham & Hartnett, 2004) from $[1990\text{-}04, t - 12\text{m}]$ by cosine similarity on the z-scored 4-D macro state vector (growth pulse, inflation pulse, term spread, VIX pulse), chosen ex ante to mirror the investment-clock dimensions, where pulses are 6-month changes and growth pulse $= -\Delta_6$(Unemployment, shift 1); *(iii)* ask the critic to compress analog evidence into one rule; *(iv)* append the rule to a rolling 6-rule buffer; *(v)* ask the actor to score seven factors in $[0, 1]$; *(vi)* evaluate the Spearman rank IC against realised cross-sectional ranks of market-adjusted residual factor returns. For every month u, we standardise the macro state vector against the expanding history available at u: $z_u = (x_u - \mu_{\leq u})/\sigma_{\leq u}$. At decision month t, cosine similarity is computed between z_t and candidate z_s for $s \leq t - 12$ months. No method uses a full-sample scaler. A representative critic rule from 2023-04 (analogs: 1990-06, 2006-04, 2006-11, 2007-01) reads: *"Given the current low inflation and rising unemployment, indicative of a growth-rising environment, raise HML by* 0.15 *and lower*

Table 1. Monthly rank IC on the 36-month walk-forward (2023-04 to 2026-03). All methods use decision-time inputs only. Macro+nowcast ridge has $n = 34$ due to two missing unemployment observations; restricting all methods to the common 34-month subset preserves the qualitative Full LLM–kNN comparison: kNN remains comparable on median IC, while Full LLM remains higher on mean IC.

Method	Mean	Median	95% CI	Perm. p
Base: zero-shot CoT	$+0.007$	$+0.027$	$[-0.15, +0.16]$	0.92
+ raw analog retrieval	$+0.006$	$+0.063$	$[-0.15, +0.16]$	0.94
+ critic rule synthesis	$+0.088$	$+0.009$	$[-0.05, +0.22]$	0.21
Full LLM: + CF nowcast	$\mathbf{+0.131}$	$+0.154$	$[-0.02, +0.28]$	$\mathbf{0.11}$
Nowcast-only ridge	$+0.045$	$+0.125$	$[-0.14, +0.23]$	0.63
Macro + nowcast ridge	$+0.058$	$+0.071$	$[-0.11, +0.23]$	0.51
kNN macro analog	$+0.070$	$\mathbf{+0.161}$	$[-0.11, +0.24]$	0.44

BAB by 0.15*, as historical analogs suggest HML tends to outperform in such macro conditions."* The rule is parsed deterministically into a Δ-vector applied to the actor's seven-factor score before final ranking.

Non-LLM baselines. We compare the LLM pipeline against three baselines drawn from the empirical asset pricing tradition (Gu et al., 2020) that consume the same structured macro / nowcast information set under the same decision-time constraint: *(i) nowcast-only ridge*: per-factor Ridge regression with $X = [\text{cf_nowcast_yoy}, \text{cf_nowcast_delta_6m}]$ only; *(ii) macro + nowcast ridge*: per-factor Ridge on the full 6-feature decision-time vector (the four pulses plus the two nowcast columns), with a `StandardScaler` fit on the training period only; *(iii) kNN macro analog*: the same $K = 4$ cosine retrieval as the LLM pipeline, but instead of an LLM critic, output the per-factor mean of the four analogs' realised next-month residual returns. Both Ridge models train only on pre-test months with complete labels and decision-time features (≈ 110 months ending 2023-03; the effective start is set by the Cleveland Fed nowcast archive coverage). The Ridge baselines are therefore conservative relative to the kNN analog baseline, which pools macro states from $[1990\text{-}04,\ t - 12\text{m}]$ with the same ≥ 12-month embargo as the LLM pipeline and does not require the nowcast feature for analog scoring. The principal apples-to-apples contrast in the results below is therefore LLM-vs-kNN, both of which draw analogs from the same 33-year history; the ridge baselines are reported as a sanity check on a smaller training horizon.

4. Results

We report three checks. *First*, does the LLM forecaster produce a positive rank signal under decision-time inputs? *Second*, is the signal stable across non-overlapping contiguous sub-windows? *Third*, does the gain come from the LLM architecture or from the real-time nowcast input?

Table 2. Three non-overlapping contiguous 12-month sub-windows of the full LLM pipeline.

Sub-window	Mean Spearman IC
2023-04 to 2024-03	$+0.111$
2024-04 to 2025-03	$+0.165$
2025-04 to 2026-03	$+0.117$

Positive rank signal, but underpowered. The full LLM pipeline (Table 1, row 4) achieves mean monthly rank IC $+0.131$ and median $+0.154$, the strongest of the four LLM ablation rows. The bootstrap 95% CI on the mean is $[-0.02, +0.28]$; the sign-symmetric permutation p is 0.11, which does not reach conventional significance.

Sign-consistent across sub-windows. Table 2 reports the mean IC of the full pipeline on three non-overlapping contiguous 12-month sub-windows; all three are positive $(+0.111, +0.165, +0.117)$.

What is doing the work? The four-row ablation in Table 1 isolates the contribution of each design element. Zero-shot Chain-of-Thought (row 1) and zero-shot plus raw analog injection into the actor's prompt (row 2) produce essentially the same mean and median IC. The critic's rule synthesis (row 3) lifts the mean to $+0.088$ but leaves the median near zero $(+0.009)$. The single largest observed change occurs when the Cleveland Fed CPI nowcast is added (row 4): the median moves from $+0.009$ to $+0.154$.

LLM versus non-LLM under the same decision-time constraint. The bottom three rows of Table 1 report ridge and kNN baselines using the same structured macro / nowcast information set under the same decision-time constraint. The kNN macro-analog baseline recovers a slightly higher median IC than the full LLM pipeline ($+0.161$ vs $+0.154$). However, the kNN baseline's mean IC $(+0.070)$ is considerably below the LLM's $(+0.131)$, and its permutation p is much higher (0.44 vs 0.11). The two ridge baselines have positive medians but mean IC below $+0.06$. We interpret this as: the macro-similar retrieval step does much of the median-IC work, while the LLM pipeline appears to improve the months in which the model commits most strongly to top and bottom factors, which are the months that drive the long-short sanity check.

Economic sanity check. Table 3 translates ranks into a simple monthly allocation: long the top two factors and short the bottom two, equal-weighted within each leg, with 5 bps cost per unit weight change. This is not a deployable trading strategy; it tests whether rank IC corresponds to non-trivial allocation value.

The LLM pipeline records the strongest Sharpe $(+0.71)$ and annualised return $(+4.2\%)$; the kNN baseline is positive but lower $(+0.29,\ +1.5\%)$; the ridge baselines lose money over the test window. The LLM-vs-kNN Sharpe gap is consistent

Table 3. Long-top-2 / short-bottom-2 sanity check, monthly rebalanced with 5 bps per unit weight change. Turnover is the average monthly L1 weight change (sum of $|w_t - w_{t-1}|$ across factors). *Not a trading claim.*

Method	Ann. Ret.	Sharpe	Max DD	Turnover
Full LLM pipeline	$+\mathbf{4.2}\%$	$+\mathbf{0.71}$	-4.7%	0.81
kNN macro analog	$+1.5\%$	$+0.29$	-5.1%	1.10
Nowcast-only ridge	-0.9%	-0.12	-10.6%	0.17
Macro + nowcast ridge	-0.2%	-0.01	-10.8%	0.82

with the mean-IC gap, suggesting that the LLM's marginal contribution is concentrated in the extreme rankings that drive long-short portfolio formation.

5. Discussion

Main finding. Our main finding is conditional and diagnostic. Under a strict decision-time information constraint, a 7B retrieval-augmented LLM produces a directionally positive factor-ranking signal on a 36-month walk-forward, with median IC $+0.154$ and positive means across three non-overlapping contiguous sub-windows. This magnitude is in the same range as monthly cross-sectional IC reported for machine-learning factor models in empirical asset pricing (Gu et al., 2020). The largest observed change in median IC occurs when the Cleveland Fed CPI nowcast is added. A simple kNN macro-analog baseline (no LLM) recovers comparable median IC but materially lower mean IC and lower long-short Sharpe. We therefore view the result as two-layered: real-time inflation information and macro-similar retrieval together account for much of the median signal, while the LLM appears to improve the extreme rankings that drive the long-short sanity check.

Why the nowcast matters. The Cleveland Fed nowcast helps because it substitutes for an unobservable quantity. At the month-end decision date for month t, the BLS-published CPI for month t does not yet exist, so any backtest that conditions on it is consuming approximately ten days of look-ahead about the realised macro state. Replacing that value with the Cleveland Fed daily nowcast restores the information set that a real decision-maker would have. This is a major and easily overlooked source of leakage we identified, and matches the broader concern that financial backtests routinely overstate performance through subtle look-ahead and selection effects (Harvey & Liu, 2020; Bailey & Lopez de Prado, 2014). The kNN baseline benefits from the same nowcast input, which is why it tracks the LLM on median IC; the LLM's residual edge concentrates in extreme cross-sectional ranks rather than in central tendency, suggesting the contribution is closer to non-linear interaction across heterogeneous factor exposures than to discovering a new signal de novo.

Limitations. *Statistical power.* The evidence is underpowered: $n = 36$ monthly decisions, mean-IC 95% CI includes zero, permutation $p = 0.11$. With this sample size and a per-month IC standard deviation near the Spearman null, even a moderately positive expected IC fails to clear conventional significance thresholds, and we explicitly avoid framing the result as a discovery; the broader cross-sectional finance literature has documented how fragile single-window evidence is once multiple-testing is taken into account (Harvey et al., 2016). *Information matching.* The LLM and kNN paths are both leakage-clean but not perfectly information-matched—the LLM critic anchors its macro narrative at $t-1$ and the actor additionally reads recent event summaries (Appendix A)—so we read LLM-vs-kNN as a comparison of logged decision protocols rather than a controlled input ablation. *Temporal dependence.* A moving-block bootstrap that accounts for autocorrelation yields mean-IC 95% CIs comparable to (and slightly tighter than) the i.i.d. interval (Appendix D.3), so the rolling design does not inflate the reported significance. *Coverage.* The evaluation covers only U.S. equity style factors and only one nowcast source. The benchmark's 2023-04 start is set by the joint availability of the frozen nowcast archive, event corpus, factor panel, and precomputed LLM decision logs under this pipeline configuration. Extending the test window would require reconstructing a longer frozen decision-time input archive and rerunning the walk-forward pipeline. Two test months are unavailable for the *macro + nowcast ridge* baseline because its full six-feature vector is incomplete in our frozen data snapshot. *Deployability.* The economic sanity check is intentionally simple and should not be interpreted as a deployable trading claim; deployable claims require accounting for transaction costs, financing, and capacity, all of which materially compress paper Sharpe ratios for academic factors (Frazzini et al., 2012).

Future work. The natural extensions are (a) longer evaluation windows once additional decision-time archive coverage is reconstructed, so power against single-window tests is no longer the binding constraint, and (b) explicit decomposition of the LLM's marginal contribution into the extremes-vs-center dimension surfaced here, which would clarify whether the mechanism is genuinely non-linear interaction or an artefact of small-n ranking variance. Both directions can be pursued without changing the leakage-controlled benchmark protocol introduced in this paper.

Reproducibility. We will release the walk-forward harness, all prompt templates, the per-month score parquets for every method in Tables 1–3, the parsed Cleveland Fed nowcast archive (frozen snapshot 2026-05-13), the retrieved analog indices, the full critic-rule log, and the exact Qwen2.5-7B-Instruct 4-bit (NF4 via bitsandbytes) inference configuration with $T = 0$ greedy decoding. All reported bootstrap (10,000 resamples) and sign-symmetric permutation tests (10,000) use `np.random.seed(42)`.

References

AQR Capital Management. AQR factor and data sets. https://www.aqr.com/Insights/Datasets, 2026. BAB and QMJ factor returns; accessed 2026-05-13.

Asness, C. S., Frazzini, A., and Pedersen, L. H. Quality minus junk. *Review of Accounting Studies*, 24(1):34–112, 2019.

Bailey, D. H. and Lopez de Prado, M. The deflated Sharpe ratio: Correcting for selection bias, backtest overfitting, and non-normality. *The Journal of Portfolio Management*, 40(5):94–107, 2014.

Chen, Z., Chen, J., Chen, J., and Sra, M. Position: Standard benchmarks fail – auditing LLM agents in finance must prioritize risk, 2025. arXiv:2502.15865.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, 2023.

Fama, E. F. and French, K. R. A five-factor asset pricing model. *Journal of Financial Economics*, 116(1):1–22, 2015.

Federal Reserve Bank of Cleveland. Inflation nowcasting. https://www.clevelandfed.org/indicators-and-data/inflation-nowcasting, 2026. Daily CPI nowcast archive; frozen snapshot accessed 2026-05-13.

Frazzini, A. and Pedersen, L. H. Betting against beta. *Journal of Financial Economics*, 111(1):1–25, 2014.

Frazzini, A., Israel, R., and Moskowitz, T. J. Trading costs of asset pricing anomalies. Fama-Miller / Chicago Booth working paper; SSRN 2294498, 2012.

French, K. R. Kenneth R. French data library. https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html, 2026. Accessed 2026-05-13.

Greetham, T. and Hartnett, M. The investment clock: Special report #1 – making money from macro. Technical report, Merrill Lynch Global Asset Allocation, November 2004. 10 November 2004. Practitioner framework; subsequent updates by BofA Merrill Lynch and Royal London Asset Management.

Gu, S., Kelly, B., and Xiu, D. Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33 (5):2223–2273, 2020.

Harvey, C. R. and Liu, Y. False (and missed) discoveries in financial economics. *The Journal of Finance*, 75(5): 2503–2553, 2020.

Harvey, C. R., Liu, Y., and Zhu, H. . . . and the cross-section of expected returns. *The Review of Financial Studies*, 29 (1):5–68, 2016.

Jegadeesh, N. and Titman, S. Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, 48(1):65–91, 1993.

Knotek, E. S. and Zaman, S. Nowcasting U.S. headline and core inflation. *Journal of Money, Credit and Banking*, 49 (5):931–968, 2017.

Laskin, M., Wang, L., Oh, J., Parisotto, E., Spencer, S., Steigerwald, R., Strouse, D. J., Hansen, S., Filos, A., Brooks, E., et al. In-context reinforcement learning with algorithm distillation. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2210.14215.

Li, H., Cao, Y., Yu, Y., Javaji, S. R., Deng, Z., He, Y., Jiang, Y., Zhu, Z., Subbalakshmi, K. P., Huang, J., Qian, L., Peng, X., Suchow, J. W., and Xie, Q. INVESTORBENCH: A benchmark for financial decision-making tasks with LLM-based agent. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2509–2525, 2025. arXiv:2412.18174.

Lopez-Lira, A. and Tang, Y. Can ChatGPT forecast stock price movements? return predictability and large language models, 2023. SSRN 4412788; arXiv:2304.07619.

Qwen Team. Qwen2.5 technical report, 2024. arXiv:2412.15115.

Shinn, N., Cassano, F., Berman, E., Gopinath, A., Narasimhan, K., and Yao, S. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Xiao, Y., Sun, E., Luo, D., and Wang, W. TradingAgents: Multi-agents LLM financial trading framework, 2024. arXiv:2412.20138.

Appendix

The appendix documents reproducibility metadata, leakage-control details, a descriptive critic-rule audit, two representative case studies, two compact sensitivity diagnostics, and extended figures. It is descriptive and intended to make the system's behaviour inspectable; it is not a causal claim about any individual factor's performance in any individual month.

A. Reproducibility and leakage controls

We will release a public reproducibility package containing the frozen input snapshots, per-method score files, retrieved analog indices, critic-rule logs, prompt templates, and the analyses needed to reproduce all tables and figures. The package includes the Cleveland Fed nowcast snapshot frozen on 2026-05-13, the Qwen2.5-7B-Instruct 4-bit NF4 inference configuration, greedy decoding with $T = 0$, and the random seeds used for bootstrap and permutation tests (`seed` = 42).

All retrieval standardisation is expanding-window. For each month u, we compute

$$z_u = \frac{x_u - \mu_{\leq u}}{\sigma_{\leq u}},$$

where $\mu_{\leq u}$ and $\sigma_{\leq u}$ are computed only from information available at or before month u. Candidate months satisfy $s \leq t-12$ months, and no method uses a full-sample scaler. We require at least 24 non-null monthly observations before computing a standardised state.

The logged LLM protocol uses an as-of anchor $t - 1$ for the critic-side macro narrative, whereas the kNN baseline anchors directly at t. Both anchors are leakage-clean because each path only consumes information available at or before its anchor. We therefore interpret LLM-vs-kNN comparisons as comparisons of the logged decision protocol, not as a controlled sensitivity to the standardisation anchor.

Prompt templates will be included in the public reproducibility package.

B. Critic-rule descriptive audit

We summarise the 36 critic outputs using deterministic keyword and pattern rules; no LLM-as-judge is used. Table 4 shows that the rules are highly concentrated in this window: of the 30 rules with extractable factor pairs, 25 raise HML, usually under inverted-yield-curve / low-volatility tags. This supports two readings. First, the LLM behaviour is inspectable rather than arbitrary. Second, the 2023–2026 window is regime-specific, so the rule distribution should not be interpreted as evidence of general factor-timing skill.

Table 4. Recurring critic-rule themes on the 36 monthly decisions of the test window. Macro regime tags are abbreviated: `iyc` = inverted yield curve, `lv` = low volatility. Total count of categorised (raise, lower) pairs = 30; the remaining 6 rules have no extractable factor pair.

Macro regime tag(s)	Raise	Lower	Count
iyc, lv	HML	BAB	7
iyc, lv	HML	UMD	6
iyc, lv	HML	CMA	5
iyc	HML	UMD	4
iyc, lv	SMB	UMD	2
iyc, lv	QMJ	BAB	1
iyc, lv	HML	RMW	1
iyc	UMD	QMJ	1
iyc	HML	QMJ	1
iyc	UMD	BAB	1
other / mixed regime tags	HML	BAB	1

C. Case studies

We present two pre-selected decision months: the best single-month Full LLM IC, and the month with the largest Full-LLM minus kNN IC gap. These examples are illustrative, not causal; they show how the logged critic rule and actor scores relate to the realised rank ordering in two months where the LLM and kNN paths diverge. The critic rules quoted below

are reproduced verbatim from the model output; any citation-like text within them (e.g., "Asness-Liew-Pedersen 2013") is generated by the model and is not an author citation.

C.1. Best LLM month: 2024-12

Decision date. 2024-12-31. **Decision-time macro snapshot.** VIX 17.4; term spread $+0.33$; HY credit spread 2.92; real 10y yield 2.24; shift-1 unemployment 4.10; shift-1 FRED CPI YoY 2.87; Cleveland Fed CPI YoY nowcast at t: $\mathbf{2.86}$.

Retrieved analogs ($K = 4$). 1990-06 ($\rho = +0.96$); 2019-03 ($\rho = +0.96$); 2023-10 ($\rho = +0.99$); 2023-12 ($\rho = +0.98$).

Critic. *Rule:* "Given the inverted yield curve and low equity volatility, raise the score of HML by 0.15 and lower the score of UMD by 0.15, as historical analogs indicate that value factors tend to outperform in such macro conditions (Asness-Liew-Pedersen 2013)." *Diagnosis:* "In all analogs, HML consistently wins while UMD consistently loses."

Actor scores vs realised ranks.

Factor	SMB	HML	RMW	CMA	UMD	BAB	QMJ
Score	0.20	0.30	0.80	0.40	0.35	0.50	0.70
Pred rank	1	2	7	4	3	5	6
Realised	1	2	7	3	4	5	6

Outcome. Full LLM IC $= +0.964$; kNN IC $= -0.321$; gap $= +1.286$.

Reading. December 2024 sits in the persistent inverted-yield-curve / low-volatility regime that characterises most of the test window. The retrieved analogs span 1990, 2019, and two 2023 months, and the critic converges on a value-tilted prior. This case is consistent with the actor recovering the realised cross-sectional ordering on five of seven factors and swapping only two adjacent middle factors (CMA, UMD); the kNN baseline mis-ranks top and bottom factors in the same month.

C.2. Largest LLM-beats-kNN gap: 2024-07

Decision date. 2024-07-31. **Decision-time macro snapshot.** VIX 16.4; term spread -0.20; HY credit spread 3.25; real 10y yield 1.85; shift-1 unemployment 4.20; shift-1 FRED CPI YoY 2.94; Cleveland Fed CPI YoY nowcast at t: $\mathbf{3.01}$.

Retrieved analogs ($K = 4$). 1990-05 ($\rho = +0.99$); 2006-03 ($\rho = +0.94$); 2006-08 ($\rho = +0.99$); 2006-09 ($\rho = +0.94$).

Critic. *Rule:* "Given the inverted yield curve and low equity volatility, raise the score of HML by 0.15 and lower the score of CMA by 0.15, as historical analogs indicate that value factors tend to outperform in such macro conditions (Asness-Liew-Pedersen 2013)." *Diagnosis:* "In all analogs, HML consistently wins while CMA and QMJ consistently lose."

Actor scores vs realised ranks.

Factor	SMB	HML	RMW	CMA	UMD	BAB	QMJ
Score	0.70	0.80	0.30	0.40	0.60	0.50	0.20
Pred rank	6	7	2	3	5	4	1
Realised	7	6	3	4	1	5	2

Outcome. Full LLM IC $= +0.607$; kNN IC $= -0.857$; gap $= +1.464$.

Reading. July 2024 has a similar macro signature to 2024-12 but a different mix of analogs: three of four come from pre-GFC 2006 episodes. The critic again produces a value-tilted prior. The actor gets the broad direction right but mis-orders UMD. This case is consistent with the critic rule acting as a coarse value-tilted abstraction of the analog evidence rather than a mechanical average of analog returns.

D. Sensitivity and extreme-rank diagnostics

D.1. Sensitivity to analog count K

We vary K only in the LLM-stripped kNN macro-analog baseline over $K \in \{2, 4, 6, 8, 12\}$, keeping the retrieval features, embargo, labels, and long-short allocation rule fixed. The LLM actor-critic pipeline is not rerun. This diagnostic tests whether the retrieval signal is specific to the $K = 4$ choice used in the main experiment.

The kNN signal is not unique to $K = 4$: several nearby choices remain positive, and $K = 8$ is strongest in this diagnostic. We retain $K = 4$ in the main comparison because it was fixed ex ante to match the LLM retrieval protocol; the sweep is reported only as a sensitivity check.

Table 5. kNN macro-analog sensitivity to the number of retrieved analogs K.

K	Mean IC	Median IC	Long-short Sharpe
2	+0.017	+0.089	+0.02
4	+0.070	+0.161	+0.29
6	+0.086	+0.089	+0.51
8	+0.133	+0.179	+0.77
12	+0.079	+0.089	+0.17

D.2. Top/bottom overlap

Because the long-short sanity check depends only on the top and bottom of the predicted factor ranking, we report top-2 and bottom-2 overlap with realised factor ranks for each decision month and average across the 36-month window. This diagnostic makes the extreme-rank interpretation directly observable; it is not a separate trading claim and does not prove LLM superiority.

We define:

$$\mathrm{Top2Overlap}_t = \frac{\left|\mathrm{Top2}_{\mathrm{pred},t} \cap \mathrm{Top2}_{\mathrm{real},t}\right|}{2}, \quad \mathrm{Bottom2Overlap}_t = \frac{\left|\mathrm{Bottom2}_{\mathrm{pred},t} \cap \mathrm{Bottom2}_{\mathrm{real},t}\right|}{2}.$$

$$\mathrm{ExtremeOverlap}_t = \tfrac{1}{2}\big(\mathrm{Top2Overlap}_t + \mathrm{Bottom2Overlap}_t\big).$$

Table 6. Extreme-rank overlap diagnostics, averaged over the 36-month walk-forward (34 months for the macro + nowcast ridge).

Method	Top-2	Bottom-2	Extreme
Full LLM pipeline	+0.417	+0.278	+0.347
kNN macro analog	+0.347	+0.361	+0.354
Nowcast-only ridge	+0.278	+0.389	+0.333
Macro + nowcast ridge	+0.309	+0.353	+0.331

The overlap diagnostic is mixed: Full LLM has the highest Top-2 overlap, whereas kNN has slightly higher Bottom-2 and average Extreme overlap. We therefore treat this table as an interpretability diagnostic rather than evidence of uniformly better extreme-rank recovery; Table 3 remains the allocation-level sanity check.

D.3. Robustness to temporal dependence

The rolling walk-forward induces temporal dependence in the monthly IC series that an i.i.d. bootstrap ignores. We recompute the full-pipeline mean-IC 95% confidence interval with a moving-block bootstrap (10,000 resamples, seed 42) at block lengths $L \in \{3, 6\}$ months. The resulting intervals are $[+0.013, +0.268]$ (L=3) and $[+0.039, +0.228]$ (L=6), comparable to and slightly tighter than the i.i.d. interval $[-0.023, +0.282]$ reported in Table 1. The monthly IC series has mildly negative lag-1 autocorrelation (-0.14), so accounting for temporal dependence does not widen the interval. We nonetheless treat the mean-IC evidence cautiously given $n = 36$ and the sign-symmetric permutation $p = 0.11$; this check addresses whether the rolling design's serial dependence inflates the reported significance.

E. Extended figures

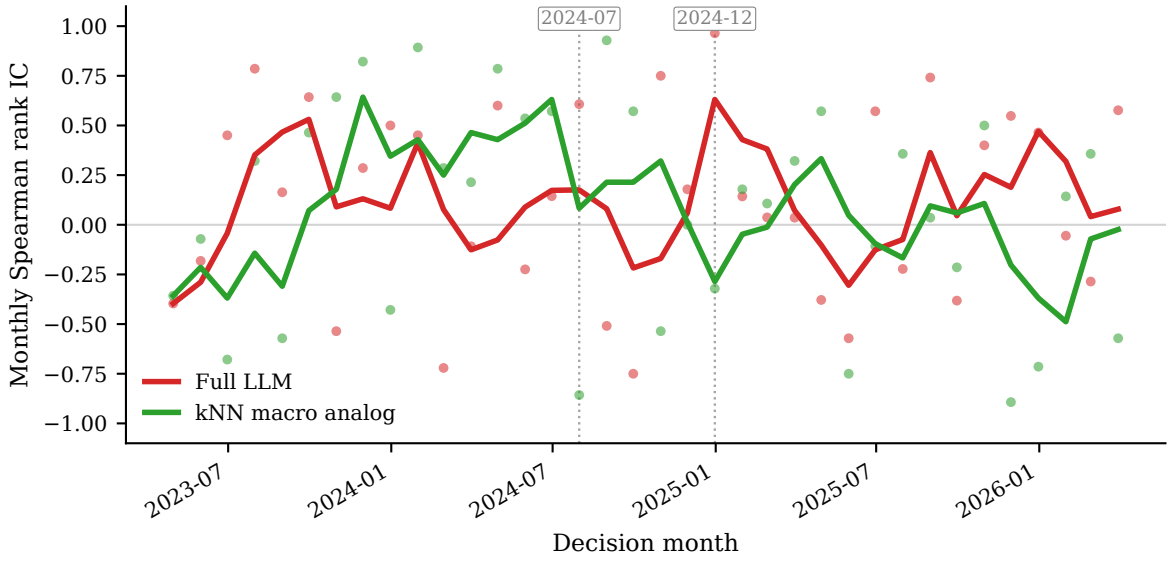

Figure 2. Monthly rank IC for the two main analog-based methods. Points show raw monthly IC; thick lines show 3-month rolling averages for readability. Summary statistics in Table 1 use unsmoothed monthly IC. Marked months correspond to the case studies in Appendix C.

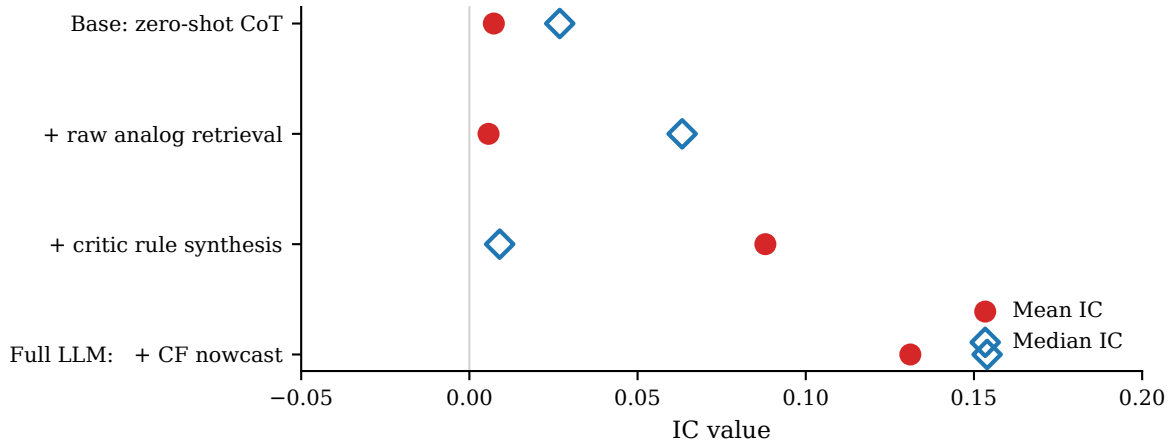

Figure 3. LLM ablation summary. Points show mean and median monthly rank IC for the four ablation rows. The largest observed median-IC change occurs when the Cleveland Fed nowcast is added.

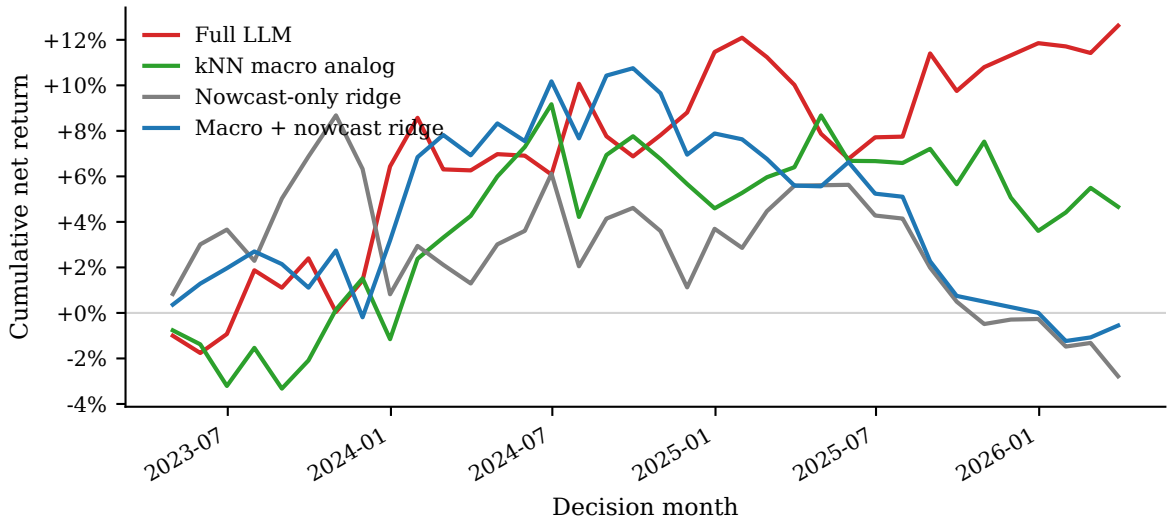

Figure 4. Cumulative net return of the long-top-2 / short-bottom-2 sanity check (monthly rebalanced, 5 bps per unit weight change), following the methodology of Table 3 of the main paper. This plot visualises the allocation sanity check only and is not evidence of deployable trading performance.