Forecasting Time-Varying Correlation Matrices with Large Language Models

Anonymous Authors[1]

Abstract

Forecasting a time-varying correlation matrix is a core financial prediction problem: portfolio variance, value-at-risk, and composite stress indices are all non-linear functions of it, and a valid forecast must stay on a constrained manifold — symmetric, unit-diagonal, positive semidefinite. We ask whether a pre-trained large language model (LLM) can do this zero-shot. We prompt Anthropic Claude to read a short history of a five-asset panel and emit the next correlation matrix directly as JSON; a light eigenvalue-repair step fixes the rare invalid output, and is needed on only $\sim$2% of calls. On two daily panels — Russia financial-stress sub-indices and Eurostoxx country returns — chain-of-thought prompting beats the persistence baseline at multi-week horizons, by a margin that grows monotonically with the horizon and reaches about $+5\%$ of the persistence-to-oracle error gap at the one-month horizon on both panels; at the one-day horizon, where persistence is near-perfect, it does not. The LLM does not beat the strongest classical baseline, exponential smoothing toward the long-run average. We explain why, and show that the LLM is not merely a worse version of that baseline: it moves the matrix in a genuinely different and informative *direction*, but takes too small a *step*. This points to hybrid forecasters that keep the LLM's direction and borrow a classical method's magnitude.

1. Introduction

Once point forecasts are fixed, many predictive tasks in finance reduce to forecasting a low-dimensional, time-varying correlation matrix C_{t+h}. Portfolio variance, value-at-risk, and composite financial-stress indices such as the CISS (Holló et al., 2012) all aggregate their inputs through a functional that is *quadratic* in the correlation forecast — so a forecast error in $\widehat{C}_{t+h}$ does not just pass through, it compounds.

Correlation forecasting is harder than ordinary univariate forecasting because the target is *constrained*: a valid correlation matrix is symmetric, has a unit diagonal, and is positive semidefinite (PSD). Forecasting each entry on its own does not, in general, produce a matrix that satisfies these constraints jointly. The classical workhorses enforce the constraint by construction. The standard parametric choice, DCC-GARCH (Engle, 2002), first filters each series' time-varying volatility with a univariate GARCH model, then evolves the correlation matrix through a parsimonious two-parameter recursion that yields a valid matrix at every step; heterogeneous-autoregressive realised-covariance models (Bollerslev et al., 2018) extend the same idea. But they are trained from scratch on numerical histories alone, and at multi-week horizons they typically beat a simple *persistence* forecast (carry C_t forward unchanged) only by a small margin.

A pre-trained large language model offers a different starting point. LLM-based time-series forecasting — LLMTime (Gruver et al., 2023), Chronos (Ansari et al., 2024), and related foundation models, which tokenise a numerical series and forecast it the way a language model predicts text — has so far targeted *univariate* series, with mixed-to-negative results on financial data (Rahimikia et al., 2025; Tan et al., 2024). In finance, LLMs have been used for return prediction (Lopez-Lira & Tang, 2023), as providers of Black–Litterman *views* that are then paired with a classically estimated covariance matrix (Lee et al., 2025), and inside neural portfolio-optimisation pipelines where the covariance is still produced by a dedicated estimator or covariance-cleaning step (Hwang et al., 2025; Bongiorno et al., 2025). In every case the covariance object itself stays classical: the LLM supplies expectations, signals, or text features, never the second-moment structure. We are unaware of prior work that asks an LLM to *directly output a correlation matrix*.

This paper asks a simple question: used zero-shot, can an LLM forecast C_{t+h}? We prompt Anthropic Claude for the matrix directly as structured JSON and project the parsed

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

output back onto the valid PSD manifold. Our contributions are:

1. **A zero-shot pipeline** (Section 2): three structured-JSON prompts paired with an eigenvalue-clip projection (Higham, 2002), binding on only ∼2% of Russia calls.
2. **A cross-panel result** (Section 3): on two daily five-asset panels — Russia financial-stress sub-indices (n=107 origins) and Eurostoxx country returns (n=72 origins) — CoT prompting beats persistence at multi-week horizons — strictly so (bootstrap CI above zero) from h=5 on Russia and at h=22 on both panels — with the margin growing monotonically with the horizon.
3. **The LLM is not a disguised shrinkage rule** (Section 3.2): a direction-magnitude decomposition shows a directional signal classical shrinkage lacks, even though the LLM moves too cautiously.
4. **A "soft-persistence" signature** (Section 4): the LLM moves only a small, quantified fraction of the realised distance — a falsifiable conjecture.

2. Methodology

Panels. Panel A (Russia CISS, n=107 origins). Five segment-level financial-stress sub-indices (money market, government bonds, equity, FX, financial intermediaries) at daily frequency, 2010-02-19 to 2026-04-29 (T=4,209). Each sub-index lies in $[0, 1]$ (a recursive empirical-CDF transform of a raw stress measure); C_t is their EWMA correlation matrix (exponentially weighted, decay λ=0.93) and a composite stress index aggregates the sub-indices and C_t through a quadratic form, the Holló–Kremer CISS formula (Section I). Our forecasting target throughout is C_t itself. **Panel B (Eurostoxx, n=72 origins).** Daily log-returns of DAX, CAC 40, FTSE MIB, IBEX 35, AEX, 2005-01-03 to 2025-10-16, evaluated directly without a downstream aggregator. Both panels share n=5 and the EWMA pipeline.

Walk-forward protocol and metric. We evaluate on an expanding window: forecast origins every 30 trading days, a minimum train window of 1,000 days, and horizons $h \in \{1, 5, 10, 22\}$. At each origin we score a forecast by its Frobenius distance to the realised matrix C_{t+h}, and compare strategies with a *paired bootstrap*: we resample the set of forecast origins with replacement B=2,000 times to obtain confidence intervals, reusing the *same* resampled origins for every strategy so all comparisons are like-for-like. Origins are 30 days apart and $h \leq 22$, so forecast windows do not overlap and an i.i.d. paired bootstrap is appropriate. We report *gap closure*: how far a strategy moves from *persistence* (carry C_t forward unchanged, scored 0%) toward the *oracle* (a forecaster handed the realised C_{t+h}, scored 100%). A positive value simply means "better than persistence". In symbols, $G_s = (\mathrm{RMSE}_{\text{pers}} - \mathrm{RMSE}_s)/(\mathrm{RMSE}_{\text{pers}} - \mathrm{RMSE}_{\text{oracle}})$.

LLM prompt designs. Three structured-JSON variants under Anthropic Claude Opus 4.7 (Anthropic, 2026): (i) *direct* (history of L=60 days plus C_t, asks for the 5×5 matrix at $t+h$ as a JSON object); (ii) *chain-of-thought* (CoT, the focal variant: prefixes a 2–4 sentence reasoning step about stress concentration and contagion before the JSON, in the style of Wei et al., 2022); (iii) *three-exemplar few-shot* (in-context (C_t, C_{t+h}) pairs at calm/moderate/stressed quantiles). One call per cell; failures fall back to persistence. Verbatim prompts in Section A. Calls use the Claude CLI with all tools disabled and a single turn (Section J): the matrix is produced purely by next-token prediction.

PSD projection and baselines. Parsed matrices are projected onto the valid-correlation manifold via a Higham-style eigen-clip (Algorithm 1, Section A); the projection is binding on only ∼2% of Russia calls (Section A). Non-LLM baselines share the same origin grid, EWMA λ and projection: *persistence* $\widehat{C}_{t+h}$=C_t; *exp-smoothing*, a *shrink-to-mean* rule $\widehat{C}_{t+h}$=$(1-\alpha_h)C_t + \alpha_h \bar{C}_{1:t}$ that pulls the current matrix a fraction α_h of the way toward the historical average matrix $\bar{C}_{1:t}$, in the spirit of covariance shrinkage (Ledoit & Wolf, 2003) (α_h grid-searched on the training prefix); DCC-GARCH (Engle, 2002); an LSTM (32 hidden units); and a look-ahead oracle. For Section 3.2 we add one *calibrated shrink-to-mean* baseline: exp-smoothing whose single shrinkage weight is scalar-tuned so that its median Frobenius step size matches the LLM's.

3. Results

3.1. Gap closure on two panels

Table 1 reports gap closure at the one-month horizon (h=22); Figure 1 traces the full horizon profile. The headline is simple: at multi-week horizons chain-of-thought prompting beats persistence, and the margin *grows monotonically* with the horizon. On Russia the gain rises monotonically — −6.4%, +2.1%, +3.0%, +5.6% at h=1, 5, 10, 22 — and is strictly positive (bootstrap CI above zero) from h=5 onward; on Eurostoxx the same monotone profile holds (−7.4%, +3.7%, +3.9%, +5.1%) and is strictly positive at h=22 (+5.1%, CI [+2.0, +9.5]). At the one-day horizon the LLM falls below persistence on both panels — as expected from the soft-persistence picture of Section 4: at h=1 persistence is near-perfect, so any move is a loss. This is the claim we defend, and no more. It does *not* beat the strongest classical baseline on the point estimate — exp-smoothing

Table 1. Gap closure (%) at the one-month horizon h=22 (matrix-Frobenius RMSE), rows sorted by Russia gap, with paired-bootstrap 95% CIs (B=2,000). Higher is better: persistence = 0 (floor), oracle = 100 (ceiling). Our method, **LLM-CoT**, is shown in bold. LLM is Claude Opus 4.7, CoT variant (direct and few-shot in Section K).

	Russia (n=107)		**Eurostoxx** (n=72)	
Strategy	Gap	95% CI	Gap	95% CI
Oracle	100	—	100	—
Exp-smoothing	+7.3	[+5.7, +8.8]	+8.5	[+3.7, +14.7]
LLM (CoT)	**+5.6**	**[+4.1, +7.0]**	**+5.1**	**[+2.0, +9.5]**
Persistence	0.0	—	0.0	—
DCC-GARCH	−12.1	[−23.2, −2.6]	−1.5	[−18.3, +15.1]

Figure 1. LLM-CoT gap closure across horizons (point estimate, 95% paired-bootstrap CI). On both panels the margin grows monotonically with the horizon: below persistence at h=1, then strictly positive (CI above zero) from h=5 on Russia and at h=22 on both panels.

reaches +7.3% on Russia and +8.5% on Eurostoxx — a gap we dissect in Section 3.2. A from-scratch LSTM fails to beat persistence on either panel and is omitted from Table 1 (Section 4).

We report only the chain-of-thought (CoT) variant in Table 1: it is the sole prompt design that parses at every origin (107/107 on Russia), whereas direct and few-shot fall back to persistence on roughly half the short-horizon origins, yielding an LLM/fallback mixture reported separately in Section K.

3.2. Is the LLM just a worse shrinkage rule?

Exponential smoothing always moves C_t a fixed fraction toward the long-run average matrix — one parameter, one direction. Since it beats the LLM on our headline metric, a natural worry is that the LLM is the same rule with more noise. We rule this out with two tests, both against a shrink-to-mean baseline scalar-tuned to the *same step size* as the median LLM step.

Magnitude. At equal step size, classical shrinkage wins: shrink-to-mean beats LLM-CoT at h=22 by +1.8 pp on Russia and +3.5 pp on Eurostoxx, both CIs excluding zero (Section B). The LLM is not under-performing merely because its step is small.

Direction. But it is not moving in the shrinkage direction. An ensemble that keeps the LLM's *direction* but rescales its step to the shrink baseline's size (Section C) closes +1.6 pp more of the gap than the raw LLM on Eurostoxx h=22 (CI [+0.45, +3.02], strictly positive) — impossible were the LLM merely noisy shrinkage. It still trails shrink-to-mean by 1.9 pp, so direction alone does not fully compensate; but the LLM is a distinct forecaster, not a degraded copy. This motivates a hybrid: keep the LLM's direction, borrow a classical method's magnitude.

4. Discussion

The LLM as a soft-persistence operator. How much does the LLM under-move the matrix? Persistence makes *no* move ($\widehat{C}_{t+h}$=C_t), the oracle the *full* realised move (C_{t+h}); the LLM nudges part of the way — a *soft-persistence operator*. We quantify the nudge by the median ratio of LLM step to realised step,

$$\rho(h) = \text{median}\, \|\widehat{C}^{\text{LLM-CoT}}_{t+h} - C_t\|_F \big/ \|C_{t+h} - C_t\|_F, \quad (1)$$

which runs from 0 (persistence) to 1 (oracle). At $h \geq 10$ we find $\rho(h)$<0.30 on both panels (Russia $\rho(22)$=0.15; Eurostoxx $\rho(10)$=0.25, $\rho(22)$=0.24). This explains the magnitude gap of Section 3.2 — the steps are simply too short — and is *falsifiable*: any daily five-asset replication can check $\rho(h)$<0.30. Candidate mechanisms are discussed in Section D.

Why the baselines land where they do. *Persistence is hard to beat*: daily correlation matrices drift rather than jump, so carrying C_t forward is already a strong forecast and the room above it is small. *Exponential smoothing* wins by exploiting their one systematic departure from a random walk — slow mean-reversion toward the long-run average. *A from-scratch LSTM* (32 units, a few thousand training days) has too little data to learn more than that drift and fails to beat persistence on either panel, so we omit it from Table 1. *DCC-GARCH* is negative at short horizons and only approaches persistence at h=22 (−12% Russia, −1.5% Eurostoxx) — the generic difficulty of a parametric correlation forecaster on a short, highly persistent sample, not a special pathology (Section E).

Honest disclosures. *Single LLM family* (Anthropic Claude); generalisation to other model families is open. *Modest n* (72–107 origins) yields wide CIs and overlapping bands between LLM-CoT and exp-smoothing on Russia; the strict claim is vs persistence. *Significance of the headline matrix-Frobenius result* rests on the paired-bootstrap CIs of Table 1; the *Diebold–Mariano* tests we report are on the secondary CISS-scalar loss, where the test vs exp-smoothing on Russia is not significant (p=0.243 squared loss; see Section G). *Pretraining-window overlap* (2005–2026) cannot be ruled out, though ρ measures displacement, not event prediction. *Multiple testing*: the h=22 headline cell stays strictly positive on both panels at Bonferroni-$\alpha/2$=2.5% (Section G). The improvements correspond to small absolute matrix-Frobenius differences ($\sim$ 0.001–0.003) and should not be interpreted as portfolio recommendations.

5. Conclusion

A pre-trained LLM can forecast a time-varying correlation matrix zero-shot: prompted for the matrix as JSON, chain-of-thought Claude beats persistence at multi-week horizons on two daily five-asset panels, by a margin that widens monotonically with the horizon to about $+5\%$ of the persistence-to-oracle gap at one month. It does not beat a calibrated classical shrinkage rule, but it is not a worse copy of one: it moves the matrix in a different, informative direction while taking too small a step — a *soft-persistence* signature. Hybrid forecasters that keep the LLM's direction and borrow a classical magnitude are the natural next step.

Future work. The one capability separating an LLM from every classical correlation forecaster is conditioning on *contemporaneous text*: a strictly no-look-ahead daily news digest in the prompt could supply the magnitude signal the pure-numerical prompt lacks (Section 4), most plausibly at regime breaks where persistence fails. Whether such a gain survives an honest information-leakage audit is the open question we leave to follow-up work.

Software and Data. Code, raw LLM responses (1,518 Claude calls), per-origin metrics and figure scripts are released anonymously with the camera-ready version (Section J). Inputs are public (MOEX, IFX, Bank of Russia, Yahoo Finance).

Impact Statement. This work characterises a new use of LLMs in finance; reported gap-closure gains are small absolute RMSE differences, not portfolio recommendations.

References

Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S. S., Arango, S. P., Kapoor, S., Zschiegner, J., Maddix, D. C., Wang, H., Mahoney, M. W., Torkkola, K., Wilson, A. G., Bohlke-Schneider, M., and Wang, Y. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024.

Anthropic. Introducing Claude Opus 4.7: Long-context (1m) frontier reasoning model. Anthropic public release notes, 2026. https://www.anthropic.com/news/claude-opus-4-7.

Bollerslev, T., Patton, A. J., and Quaedvlieg, R. Modeling and forecasting (un)reliable realized covariances for more reliable financial decisions. *Journal of Econometrics*, 207 (1):71–91, 2018.

Bongiorno, C., Manolakis, E., and Mantegna, R. N. End-to-end large portfolio optimization for variance minimization with neural networks through covariance cleaning. *arXiv preprint arXiv:2507.01918*, 2025.

Diebold, F. X. and Mariano, R. S. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13 (3):253–263, 1995.

Engle, R. Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business & Economic Statistics*, 20(3):339–350, 2002.

Gruver, N., Finzi, M., Qiu, S., and Wilson, A. G. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems*, 36, 2023.

Harvey, D., Leybourne, S., and Newbold, P. Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, 13(2):281–291, 1997.

Higham, N. J. Computing the nearest correlation matrix—A problem from finance. *IMA Journal of Numerical Analysis*, 22(3):329–343, 2002.

Holló, D., Kremer, M., and Lo Duca, M. CISS – A composite indicator of systemic stress in the financial system. Working Paper 1426, European Central Bank, 2012.

Hwang, Y., Kong, Y., Zohren, S., and Lee, Y. Decision-informed neural networks with large language model integration for portfolio optimization. *arXiv preprint arXiv:2502.00828*, 2025.

Ledoit, O. and Wolf, M. Honey, I shrunk the sample covariance matrix. *Journal of Portfolio Management*, 30(4): 110–119, 2003.

Lee, Y., Kim, Y., Kim, J., Kim, S., and Lee, Y. LLM-enhanced black–litterman portfolio optimization. *arXiv preprint arXiv:2504.14345*, 2025. ICLR 2025 Workshop on Advances in Financial AI.

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

Rahimikia, E., Ni, H., and Wang, W. Re(Visiting) time series foundation models in finance. *arXiv preprint arXiv:2511.18578*, 2025.

Tan, M., Merrill, M. A., Gupta, V., Althoff, T., and Hartvigsen, T. Are LLMs useful time-series forecasters? *arXiv preprint arXiv:2406.16964*, 2024.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

A. Verbatim prompt templates and PSD projection

A.1. Eigen-clip projection algorithm

Algorithm 1 Eigen-clip projection (Higham, 2002)

Input: candidate $C \in \mathbb{R}^{n \times n}$.
Symmetrise: $C \leftarrow \frac{1}{2}(C + C^\top)$.
Set diagonal: $C_{ii} \leftarrow 1$ for all i.
Eigendecompose: $C = V \operatorname{diag}(\lambda) V^\top$.
Clip negative eigenvalues: $\lambda_i \leftarrow \max(\lambda_i, 10^{-6})$.
Reconstruct: $C \leftarrow V \operatorname{diag}(\lambda) V^\top$.
Renormalise: $C \leftarrow D^{-1/2} C D^{-1/2}$, $D = \operatorname{diag}(C)$.
Clip off-diagonals: $C_{ij} \leftarrow \operatorname{clip}(C_{ij}, -0.999, 0.999)$ for $i \neq j$.
Symmetrise once more. Return C.

We empirically verify that post-projection minimum eigenvalues remain $\geq 10^{-6}$ on all 1,518 successful Claude calls. Across the 265 Russia calls for which we recorded the projection delta, only $\sim$2% have a binding projection (projected-vs-raw Frobenius distance exceeding 10^{-3}); in a separate eigenvalue-clip diagnostic over 751 calls, 1.3% trigger eigen-clipping.

A.2. Direct variant

```
You will receive a daily history of 5 financial sub-indices, each
bounded in [0,1], for the last L=60 trading days, plus the current
5x5 EWMA correlation matrix. Predict the 5x5 correlation matrix
that will hold h trading days from now.

History (oldest -> newest, columns = money, bond, equity, fx, fin):
{HISTORY_TABLE}

Current EWMA correlation matrix:
{C_T_MATRIX}

Return ONLY JSON in this exact format, no commentary, no code fences:
{"correlation_matrix": [[1.0, c12, c13, c14, c15],
                        [c12, 1.0, c23, c24, c25],
                        [c13, c23, 1.0, c34, c35],
                        [c14, c24, c34, 1.0, c45],
                        [c15, c25, c35, c45, 1.0]]}
```

A.3. Chain-of-thought variant

```
[Same setup as Direct, plus:]

Before outputting the JSON, briefly reason in 2-4 sentences (max 80
words) about: (i) which segments are currently most stressed; (ii)
whether stress is concentrated in one segment or spreading; (iii)
whether you expect correlations to rise (stress contagion) or fall
(idiosyncratic noise). Then output the JSON on a separate final line.
```

A.4. Few-shot variant

```
[Same setup as Direct, with three prepended examples drawn from the
training period at CISS quantiles 0.10 (calm), 0.40 (moderate),
0.80 (stressed):]
```

```
Example 1 (calm regime, t = YYYY-MM-DD):
  Input C_t = [...]
  h trading days later, C_{t+h} = [...]

Example 2 (moderate regime, ...):
  ...

Example 3 (stressed regime, ...):
  ...

Now predict C_{t+h} for the current case.
```

B. Calibrated shrink-to-mean baseline

To test whether the LLM is simply a noisier version of exponential smoothing, we compare it against a *calibrated shrink-to-mean* baseline: exp-smoothing $\widehat{C}_{t+h}{=}(1-\alpha)C_t+\alpha\bar{C}_{1:t}$ whose single weight α is scalar-tuned so that its median Frobenius displacement from C_t matches the LLM's. This equalises *step size* across the two forecasters, so any performance difference is attributable to *direction* alone. Table 2 reports the result at $h{=}22$: on both panels the calibrated shrink-to-mean beats LLM-CoT, with a paired-bootstrap difference excluding zero — at equal step size, moving straight toward the long-run mean is the better move. This *calibrated* shrink-to-mean — a single scalar weight tuned to match the LLM's median step size — is a distinct forecaster from the per-horizon grid-searched exp-smoothing of Table 1, even though their $h{=}22$ gap closures round to similar values ($+7.4$ vs $+7.3$ on Russia; both $+8.9$ on Eurostoxx). Section 3.2 and Section C show, however, that the LLM's *direction* still carries signal the shrinkage rule lacks.

Table 2. Calibrated shrink-to-mean versus LLM-CoT at $h{=}22$ under matrix-Frobenius RMSE. The shrink-to-mean gap closure is the paired-bootstrap mean; the LLM-CoT row repeats the point estimate of Table 1. Gap CI: gap closure vs persistence. Diff-vs-CoT: paired-bootstrap difference $G_s - G_{\text{LLM-CoT}}$.

Strategy	Gap (%)	Gap CI	Diff vs CoT (pp)	Diff CI (pp)
Russia, $h{=}22$ ($n{=}107$):				
shrink-to-mean	$+7.4$	$[+5.8, +8.9]$	$+1.80$	$[+0.23, +3.27]$
LLM-CoT	$+5.6$	$[+4.1, +7.0]$	–	–
Eurostoxx, $h{=}22$ ($n{=}72$):				
shrink-to-mean	$+8.9$	$[+3.7, +14.7]$	$+3.51$	$[+1.03, +6.21]$
LLM-CoT	$+5.1$	$[+2.0, +9.5]$	–	–

C. Direction-magnitude ensemble

If LLM forecasts and shrink-to-mean both attain comparable Frobenius displacement on Eurostoxx $h{=}22$, do they encode different *directions* on the constraint manifold? We test this by an ensemble that takes the LLM displacement direction and rescales it to shrink-to-mean's magnitude:

$$\widehat{C}^{\text{ens}}_{t+h} = C_t + \frac{\|d^{\text{shrink}}\|_F}{\|d^{\text{LLM}}\|_F}\, d^{\text{LLM}}, \tag{2}$$

where $d^{\text{LLM}} = \widehat{C}^{\text{LLM}}_{t+h} - C_t$ and $d^{\text{shrink}} = \widehat{C}^{\text{shrink}}_{t+h} - C_t$; the result is PSD-projected (Algorithm 1). On Eurostoxx $h{=}22$ the ensemble closes $+7.0\%$ (paired-bootstrap mean; CI $[+2.6, +12.1]$), with ensemble $-$ pure LLM $= +1.6$ pp (CI $[+0.45, +3.02]$ pp, strictly positive) and ensemble $-$ shrink-to-mean $= -1.9$ pp (CI $[-4.08, -0.06]$ pp, strictly negative). The LLM provides usable directional information; its understatement of magnitude is the dominant lag source.

D. Soft-persistence ratio: numerical table and mechanisms

Table 3 reports the median ratio $\rho(h) = \text{median}\, \|\widehat{C}^{\text{LLM-CoT}}_{t+h} - C_t\|_F / \|C_{t+h} - C_t\|_F$ across both panels.

Table 3. Soft-persistence ratio $\rho(h)$ for LLM-CoT (median across origins). Sources: `output/llm_corr_persistence_decomp_summary.csv` (Russia); `output/llm_corr_softpers_eurostoxx.csv` (Eurostoxx). Russia h=10 was not computed.

Panel / horizon	h=1	h=5	h=10	h=22
Russia	0.29	0.19	–	0.15
Eurostoxx	0.56	0.33	0.25	0.24

Falsifiable conjecture. *On any daily five-asset panel where a pre-trained LLM is prompted with the current EWMA correlation matrix and asked to forecast h days ahead in zero-shot, the median LLM-to-oracle Frobenius-displacement ratio satisfies $\rho(h)<0.30$ at $h \geq 10$.* The exact value is panel-dependent; we propose the inequality bound as a single-scalar prediction that future replications can directly check.

Three candidate mechanisms. (i) *Training-data locality*: numerical correlation matrices are rare in pretraining corpora, so the posterior over $\widehat{C}_{t+h}$ is sharply peaked at the input C_t. (ii) *Output-token cautiousness*: alignment biases against extreme numerical outputs, favouring small displacements. (iii) *JSON-format autoregression*: committing to 15 unique off-diagonal entries in one decoder pass, the maximum-likelihood completion given a strong anchor is a small perturbation.

E. DCC-GARCH: input-consistency diagnosis

DCC-GARCH must be fed inputs consistent with the evaluation target. On the Eurostoxx panel the pipeline is consistent by construction: DCC is fit on the same log-returns whose EWMA correlation is the target, and DCC there reaches -1.5% gap closure at h=22 (CI $[-18.3, +15.1]$, indistinguishable from persistence). An earlier Russia run was *not* consistent: it fit GARCH+DCC on the first differences of the logit-transformed sub-indices and therefore forecast the conditional correlation of sub-index *innovations*, while the target C_{t+h} is the EWMA correlation of the sub-index *levels*. Scoring a difference-correlation against a level-correlation is an apples-to-oranges comparison and accounts for roughly half of the apparent failure: that setup gave -26.4% at h=22.

Re-running Russia DCC on the same footing as Eurostoxx — fitting GARCH+DCC on the expanding-mean-centred sub-index *levels* that define the EWMA target — raises the h=22 gap closure to -12.1% (CI $[-23.2, -2.6]$), the value reported in Table 1. With this correction the two panels share one horizon profile: DCC is strongly negative at short horizons on *both* (Russia $-577/-117/-12\%$, Eurostoxx $-177/-39/-1.5\%$ at h=1/5/22) and only approaches persistence at the one-month horizon. The mechanism is generic: DCC's forecast reverts toward an unconditional correlation $\bar{Q}$ and uses a GARCH-filtered correlation estimator that differs from the EWMA target, so at short horizons — where persistence is near-perfect — any deviation is a loss. DCC remains below persistence on Russia at h=22, but the residual gap is the difficulty of a from-scratch parametric forecaster on a short, regime-switching sample, not a pathology specific to bounded CDF-transformed composites.

F. CISS-aggregate scalar metric on Russia (secondary)

Russia CISS is a non-linear aggregator: $\text{CISS}_t = \sqrt{w^\top (C_t \odot s_t s_t^\top) w}$. Plugging an LLM correlation forecast into this Holló-Kremer formula with cached ARIMA sub-index forecasts produces a scalar CISS forecast; bootstrapping this scalar against the realised CISS gives a different metric than the matrix-Frobenius RMSE used in Table 1. On Russia at h=5, LLM-CoT closes $+28.8\%$ of the persistence-to-oracle CISS-scalar gap (paired-bootstrap CI $[+7.3, +56.0]\%$; `output/llm_corr_bootstrap.csv`). This number is aggregator-specific and not directly comparable to Eurostoxx (no downstream non-linear aggregator); we therefore do not present it as a primary headline. The matrix-Frobenius result in Table 1 is the canonical Russia correlation finding.

G. Diebold–Mariano and multiple-testing audit

DM tests on Russia CISS-scalar loss. Diebold–Mariano tests (Diebold & Mariano, 1995) with the Harvey–Leybourne–Newbold small-sample correction (Harvey et al., 1997) (`output/llm_corr_dm.csv`) at Russia h=5: LLM-CoT vs persistence is significant (p=0.036 absolute, p=0.061 squared); LLM-CoT vs exp-smoothing is *not* significant (p=0.243 squared, p=0.467 absolute). The strictly-positive bootstrap CI for LLM-CoT on Russia is therefore a statement vs

persistence; the comparison vs exp-smoothing is inconclusive at n=107. We do not run DM on the matrix-Frobenius loss; significance on that headline metric is established by the paired-bootstrap CIs of Table 1.

Multiple-testing. We examine (2 panels) × (3 horizons) × (several strategies) cells; the panel-specific peak-horizon decision is post-hoc. Under Bonferroni-$\alpha/(2 \times 3)$=0.83% the Russia h=22 CI $[+4.1, +7.0]$ remains strictly positive while Eurostoxx h=22 shrinks closer to zero; at the chosen h=22 headline cell alone (Bonferroni-$\alpha/2$=2.5% across two panels), both CIs remain strictly positive.

Baseline-tuning protocol. The exp-smoothing α grid-search runs inside the walk-forward loop on the same origin grid, so the chosen α uses a small amount of test-set information. Since this advantages exp-smoothing (not LLM-CoT), the leakage strengthens our headline.

H. ECB U2 weekly contributions: null result

We also evaluated LLM-CoT on the European Central Bank's publicly distributed CISS sub-index contributions for the euro area aggregate (Holló et al., 2012). These are published only at weekly frequency, and four of five contributions are weekly-aggregated from underlying daily measures. Walk-forward evaluation on $n \approx 75$ origins (1999–2025) yielded LLM-CoT gap closure of $+0.3\%$ at h=5 weeks (CI $[-25.2, +29.6]\%$) and similar nulls at $h \in \{1, 22\}$ weeks. We attribute the null to two compounding effects: weekly aggregation removes the high-frequency signal that LLM-CoT captures on daily data, and ECB contributions have $\sim 5\times$ less dynamic range than the underlying raw sub-indices used in Panel A. We treat this as evidence that the methodology generalises across markets only when matched at the level of daily raw sub-indices, not at the level of pre-aggregated weekly contributions.

I. Russia CISS construction details

Russia CISS is a five-segment daily index following Holló et al. (2012) with Russia-specific raw measures. Money market: absolute deviation of RUONIA from its rolling 60-day mean. Government bonds: 22-day realised volatility of the 10-year OFZ yield. Equity: 22-day realised volatility of the IMOEX index. FX: 22-day realised volatility of the USD/RUB exchange rate. Financial intermediaries: 22-day realised volatility of the MOEX financial sector index. Each raw measure is converted to $[0, 1]$ via a recursive empirical CDF on the expanding history. Sub-indices are aggregated via the Holló-Kremer portfolio formula with equal weights. Full series spans 4,209 trading days, 2010-02-19 to 2026-04-29.

J. Reproducibility and cost

All experiments ran on a single CPU laptop with 8 GB RAM. Each forecast is one `claude -p` invocation issued with tool use disabled (`--allowed-tools ""`), a single-turn cap (`--max-turns 1`) and no session persistence. The model therefore cannot execute code, read files, browse, or iterate across turns: it emits exactly one text completion, which we parse directly. Total Claude CLI call attempts across rounds, as recorded in the raw response CSVs (`output/llm_corr_raw*.csv`):

- Russia (all variants, all horizons): 1,146 attempts / 1,080 successful (66 fallback persistence).
- Eurostoxx CoT: 288 attempts / 223 successful.
- Eurostoxx CoT fill round: 65 attempts / 65 successful.
- ECB CoT (Appendix H null result): 150 attempts / 150 successful.

Total: 1,649 **attempts,** 1,518 **successful.** Median latency 4–8 seconds per call (CoT slowest); 100% valid-correlation-matrix parse rate on the 1,518 successful calls after Higham eigen-clip projection (Algorithm 1). Bootstrap analysis with B=2,000 takes under one minute per panel. No GPU is required.

LLM-direct and LLM-fewshot success-rate disclosure. Of the Russia direct-prompt attempts, only 54/107 origins yielded successfully parsed JSON matrices at $h \in \{1, 5, 22\}$ (`output/llm_corr_forecasts_direct.npz`, `success_flags`); the remaining 53 origins per horizon hold persistence as a fallback. Few-shot has 54/107 success at h=1, 106/107 at h=5, 107/107 at h=22. CoT has 107/107 at all three horizons. The main-table result on LLM-CoT

is therefore on the pure successful-LLM sample; the LLM-direct and LLM-fewshot variants are reported separately in Section K together with the mixture caveat.

K. LLM variant results (direct, few-shot)

Table 4 reports the bootstrap gap closure for LLM-direct and LLM-fewshot on the Russia panel under the CISS-aggregate scalar bootstrap (the only bootstrap that includes these variants; `output/llm_corr_bootstrap.csv`). The headline caveat: the $n{=}107$ bootstrap for LLM-direct at $h{\in}\{1, 5, 22\}$ and LLM-fewshot at $h{=}1$ is on a *mixture sample* of 54 real LLM forecasts and 53 persistence-fallback origins; we have not re-bootstrapped on the $n{=}54$ success subsample. The reported CIs should therefore not be interpreted as pure LLM-direct or LLM-fewshot performance.

Table 4. Russia CISS-aggregate scalar gap closure for LLM-direct and LLM-fewshot, with success-rate disclosure. CIs are paired-bootstrap 95% on $n{=}107$ origins, mixing real LLM cells with persistence-fallback cells where indicated.

Variant	h	Successes	Gap (%)	CI lo	CI hi
LLM-direct	1	54/107	$+7.6$	-21.3	$+62.7$
LLM-direct	5	54/107	$+5.2$	$+0.2$	$+11.2$
LLM-direct	22	54/107	$+0.4$	-2.5	$+2.9$
LLM-fewshot	1	54/107	$+8.8$	-0.6	$+74.0$
LLM-fewshot	5	106/107	$+13.4$	$+0.3$	$+28.6$
LLM-fewshot	22	107/107	$+9.6$	-2.4	$+24.9$

The $h{=}5$ LLM-direct CI $[+0.2, +11.2]\%$ is technically strictly positive but conditional on the 50/50 LLM/fallback mixture; the strictly-positive lower bound of $+0.2$ is dominated by the 54 real LLM cells while the 53 zero-gap fallback cells artificially shrink the variance. We do not interpret this as a pure-LLM-direct positive result.