CoRE-TSF: When to Trust Time-Series Forecasts via Cross-Modal Consistency

Anonymous authors
Paper under double-blind review

Abstract

Aggregate forecast accuracy does not reveal whether a particular forecast can be trusted. Yet black-box point forecasters may expose only their numerical predictions, providing no direct signal of which individual forecasts are likely to fail. We introduce **CoRE-TSF**, an output-level framework that cross-checks each numerical forecast against an LLM's qualitative expectation of future change. Given the same observed history, a textual pathway predicts the direction and magnitude of future change from categorical history descriptors, while a numerical pathway forecasts future values. Neither pathway observes the other's output. CoRE-TSF maps both outputs to a shared ordinal space and uses their disagreement to compute a forecast-level risk score without requiring realized future values, historical forecast-error labels, or access to forecaster internals. Our primary method, CoRE-Ensemble, averages risk scores from four LLM agents before ranking forecasts. We evaluate five time-series foundation forecasters across 22 dataset families, with two LLM-based forecasters evaluated on a separate matched short-horizon cohort. On the primary cohort, CoRE-Ensemble achieves a hierarchically macro-averaged within-configuration Spearman correlation of $\boldsymbol{\rho_{\mathrm{macro}} = 0.396}$ and normalized selective utility of $\boldsymbol{U_{\mathrm{macro}} = 0.420}$, where U measures the fraction of improvement over random retention attainable under ideal error-based ordering. These results support cross-modal consistency as an externally observable signal for ranking forecast reliability.

1 Introduction

Time-series forecasting progressed from classical statistical methods, including ARIMA (Box et al., 2015) and exponential smoothing (Hyndman et al., 2008), to probabilistic autoregressive models, basis-expansion networks, and attention-based and patch-based Transformers (Salinas et al., 2019; Oreshkin et al., 2020; Lim et al., 2020; Nie et al., 2023). Recent work incorporates language models by representing time series as text or token sequences (Gruver et al., 2023; Xue and Salim, 2023) or aligning temporal representations with pretrained language-model embeddings (Jin et al., 2023).

Despite these advances, most LLM-based forecasting research focuses on predictive accuracy rather than determining when an individual forecast can be trusted. Conformal prediction and its time-series extensions address predictive coverage through prediction sets (Vovk et al., 2005; Romano et al., 2019; Xu and Xie, 2023; Gibbs and Candes, 2021). We address a complementary question: how to rank the reliability of black-box point forecasts without calibration data or historical forecast-error labels, particularly when model internals or predictive probabilities are unavailable.

Existing LLM confidence methods, including token probabilities, self-evaluation (Kadavath et al., 2022), verbalized uncertainty (Lin et al., 2022), and self-consistency (Wang et al., 2023), derive reliability signals from the same model or generation process. Related work studies cross-modal consistency (Zhang et al., 2024) and selective prediction (Geifman and El-Yaniv, 2019; Srinivasan et al., 2024). We investigate whether agreement between a separately generated textual outlook and a numerical forecast can serve as a reliability signal. Figure 1(a) illustrates the challenge: a point forecast provides a prediction, but no direct indication of its reliability.

We propose **CoRE-TSF** (*Consistency-based Reliability Estimation for Time-Series Forecasting*), a black-box framework with two procedurally separate pathways operating on the same observed history. Agent A generates a textual outlook from categorical history descriptors, while the forecasting

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

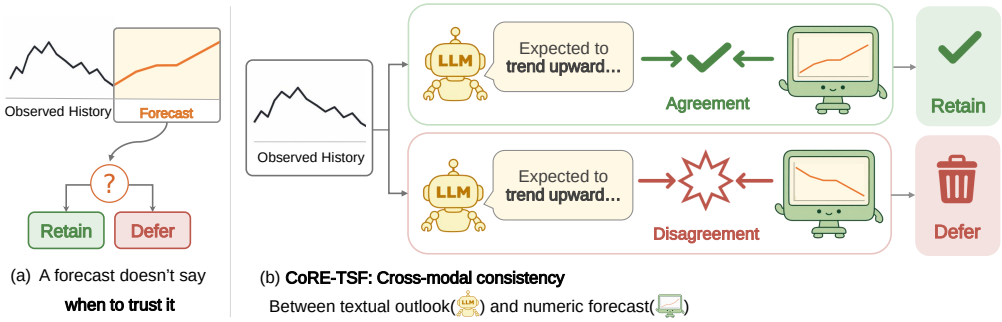


Figure 1: **Motivation and overview of CoRE-TSF.** (a) A forecast alone does not indicate when to trust it. (b) CoRE-TSF compares an LLM's textual outlook with a numeric forecast, using cross-modal consistency as a reliability signal for retain/defer decisions without guaranteeing accuracy.

model produces a numerical prediction. Neither pathway observes the other's output. CoRE-TSF maps their assessments of future direction and magnitude to a shared ordinal space, yielding consistency C_{mag} and risk $C_{\text{risk}} = 1 - C_{\text{mag}}$. Greater disagreement indicates higher estimated risk (Figure 1(b)); the score provides an ordering, not a calibrated failure probability.

Our primary method, *CoRE-Ensemble*, averages risks from four Agent A models before ranking windows. It uses equal weights without fitting to observed errors or modifying the numerical forecast. Single-agent CoRE assesses sensitivity to the choice of Agent A.

We evaluate risk–error association and selective forecasting across 22 dataset families and five backbones, with LLM-based forecasters evaluated on a separate matched short-horizon cohort.

Lower risk is associated with lower realized error on average across the evaluated forecasting models and Agent A models. On the primary cohort, CoRE-Ensemble achieves $\rho_{\text{macro}} = 0.396$ and $U_{\text{macro}} = 0.420$, supporting selection relative to random retention. Benefits vary across families, configurations, and coverage levels, without uniform superiority over alternative signals.

The contributions of this work are:

- We propose **CoRE-TSF**, an output-level framework that ranks individual forecasts by textual–numerical consistency without modifying the forecasts or requiring forecast-model internals or historical forecast-error labels.
- We establish positive within-configuration risk–error association on average across multiple forecasting models and Agent A models under hierarchical macro-averaging.
- We demonstrate selective utility relative to random retention and characterize comparative benefits and variation across dataset families and coverage levels.

2 Related Work

Language Models and Foundation Models for Time-Series Forecasting Recent foundation-model approaches to forecasting follow two paths. Methods such as LLMTime, PromptCast, and LSTPrompt recast numerical forecasting through language or prompting interfaces (Gruver et al., 2023; Xue and Salim, 2023; Liu et al., 2024), while time-series foundation models such as Chronos, TimesFM, Moirai, Time-MoE, and TiRex pretrain directly on large temporal corpora (Ansari et al., 2024; Das et al., 2024; Woo et al., 2024; Shi et al., 2025; Auer et al., 2025). Despite these advances, evaluations report sensitivity to input perturbations and question pretrained language components' contribution to some LLM-based forecasters (Park et al., 2025; Tan et al., 2024).

A related line of work integrates textual and temporal representations more directly. Time-LLM reprograms temporal patches through text prototypes (Jin et al., 2023), while GPT4MTS, TimeCMA, BALM-TSF, CALF, and CC-Time combine or align language-derived and numerical representations (Jia et al., 2024; Liu et al., 2025a; Zhou et al., 2025; Liu et al., 2025c; Chen et al., 2025). These methods primarily improve forecasting or representation learning. CALF's output-consistency loss

aligns textual and temporal branches during supervised training rather than diagnosing a completed forecast's reliability (Liu et al., 2025c). In contrast, CoRE-TSF leaves the forecaster unchanged and uses disagreement between separately generated textual and numerical outputs to rank the risk of individual forecasts.

Uncertainty Quantification and Selective Time-Series Forecasting Probabilistic forecasters characterize uncertainty through predictive distributions, while ensemble methods use dispersion across independently trained numerical predictors as an uncertainty signal (Gneiting and Katzfuss, 2014; Lakshminarayanan et al., 2017). Conformal methods further provide prediction sets or intervals with coverage guarantees, including conformalized quantile regression and extensions for temporal dependence and distribution shift (Romano et al., 2019; Stankeviciute et al., 2021; Xu and Xie, 2023; Gibbs and Candes, 2021). Although conformal scores or interval widths may provide instance-level uncertainty information, their formal objective is calibrated predictive coverage rather than ranking realized point-forecast error.

Selective prediction instead trades coverage against predictive error (Geifman and El-Yaniv, 2019). In time-series forecasting, Time-Energy learns an energy-based confidence score, while subsequent methods derive selection signals from predicted error, ambiguity or novelty, historical failure patterns, local shapelets, and horizon-specific abstention mechanisms (Brusokas et al., 2025; 2026; Liu et al., 2025d; Feng et al., 2026; Tomar et al., 2026; Inácio et al., 2026; Stradiotti et al., 2026). CoRE-TSF studies a complementary signal: disagreement between a textual assessment of future change and a numerical point forecast. Its score requires neither future observations nor access to forecaster internals at scoring time, and no selector is trained on forecast-error labels. Accordingly, it provides an empirical risk ranking rather than a calibrated error probability or coverage guarantee.

Confidence and Consistency-Based Reliability in Language Models Language-model confidence has been estimated from self-evaluation, verbalized probabilities, and token-level likelihoods (Kadavath et al., 2022; Lin et al., 2022; Tian et al., 2023). Repeated-generation methods use agreement across sampled outputs to improve decoding or quantify uncertainty (Wang et al., 2023; Manakul et al., 2023; Farquhar et al., 2024). SAC3 and CrossCheckGPT extend cross-checking to semantically equivalent prompts or peer models, while multimodal LLM evaluation also examines cross-modal consistency (Zhang et al., 2023; Sun et al., 2024; Zhang et al., 2024).

In forecasting, ReC4TS generates and aggregates multiple reasoning or forecast paths to evaluate self-consistency and improve accuracy, while MemCast selects reasoning trajectories by semantic consistency with retrieved, previously validated forecasting experience (Liu et al., 2025b; Tao et al., 2026). To our knowledge, prior work has not systematically tested whether textual–numerical agreement about future change ranks realized forecast error across heterogeneous forecasters. CoRE-TSF maps procedurally separate outputs into a shared direction–magnitude space and uses disagreement as a selective risk signal, without assuming statistical independence.

3 METHOD

3.1 PROBLEM FORMULATION

Given an observed time series $\mathbf{x}_{1:T} = (x_1, \ldots, x_T)$, our goal is to assign a risk score to an individual point forecast before the future target $\mathbf{y}_{T+1:T+H}$ is observed, without requiring forecast-model internals, predictive probabilities, or a separately trained error predictor.

CoRE-TSF constructs two procedurally separate outputs from the same observed history:

$$\mathbf{r} = f_A(\phi(\mathbf{x}_{1:T})), \qquad \hat{\mathbf{y}}_{T+1:T+H} = f_B(\mathbf{x}_{1:T}),$$

where ϕ encodes the history as categorical structural descriptors, Agent A (f_A) produces qualitative future-change assessments $\mathbf{r} = (r_1, \ldots, r_K)$, and Agent B ($f_B$) produces the numerical point forecast $\hat{\mathbf{y}}_{T+1:T+H}$. CoRE-TSF maps these outputs to a shared seven-level ordinal space and compares them over ordered future segments to obtain $C_{\text{risk}} \in [0, 1]$, with larger values indicating greater cross-modal disagreement.

Procedural separation means that neither generation path receives the intermediate or final output of the other; it implies neither statistical independence nor different model families.

3.2 TEXTUAL FUTURE-CHANGE PATH

The primary experiments use a fixed input representation and prompt. The observed history is divided into $K = 3$ contiguous segments using a deterministic equal-split rule. Within each segment, direction and magnitude are categorized from the signed and absolute ordinary least-squares (OLS) slope over integer time steps, respectively; volatility is categorized from the standard deviation of the segment values. These descriptors use fixed thresholds on unnormalized statistics and are therefore sensitive to the series' measurement scale.

Agent A receives only these descriptors, not raw values, Agent B's forecast, forecast horizon, sampling frequency, timestamps, or future segment lengths. It produces one qualitative sentence for each of K ordered relative future segments, without explicit time spans, as `Segment 1: ...`, ..., `Segment` K`: ...`. Numerical outputs and task explanations are prohibited.

After deterministic output-format normalization, each parsed r_i maps to an ordinal text-side signal:

$$\mathcal{S} = \{-3, -2, -1, 0, 1, 2, 3\}, \qquad s_{\text{text}}^{(i)} = \arg\max_{\ell \in \mathcal{S}} \max_{a \in \mathcal{A}_\ell} \cos(g(r_i), g(a)) \,.$$

where g is the frozen BGE sentence encoder (Xiao et al., 2024) and $\mathcal{A}_\ell$ is the fixed semantic anchor set for level ℓ. The sign encodes direction, zero denotes neutral change, and absolute levels 1, 2, and 3 encode weak, moderate, and strong change, respectively (e.g., upward anchors include "inch upward," "moderate growth," and "surge sharply"). These are semantic categories, not textual estimates of numerical rates. The primary results use BGE mapping without keyword fallback; keyword scores stored during generation are replaced by BGE scores for evaluation. The full prompt, descriptor thresholds, anchors, and parsing rules are provided in Appendix B.

3.3 NUMERICAL FUTURE-CHANGE PATH

The numerical path accepts Agent B's point forecast without modifying the model or accessing its internal states. The H forecast values are partitioned into K ordered, contiguous segments using a deterministic equal-split rule, with remainder handling specified in Appendix B. Correspondence with Agent A's future segments is by index, rather than explicit time spans supplied to Agent A.

For segment i, let ℓ_i denote its length, $\bar{y}_i$ its mean forecast value, and $\hat{m}_i$ its OLS slope over integer time steps. For $\ell_i \geq 2$, we compute

$$z_i = \frac{\hat{m}_i}{|\bar{y}_i| + \epsilon}, \qquad \epsilon = 10^{-8}, \qquad s_{\text{val}}^{(i)} = Q(z_i) \in \mathcal{S}.$$

Thus, z_i measures relative change per time step, not total change over the segment. The fixed quantizer Q maps it to the same ordinal direction–magnitude levels, with a dead zone around zero; exact thresholds and boundary conventions are given in Appendix B. Near-zero means can amplify $|z_i|$, and additive offsets change its normalization. Its physical-time interpretation depends on sampling frequency, while equal values need not imply equal total changes across segment lengths.

Segments with $\ell_i < 2$ have no estimable slope and are excluded from the cross-modal average. Figure 2 summarizes the procedure.

3.4 CROSS-MODAL CONSISTENCY AND RISK

Let $\mathcal{V}$ denote the nondegenerate forecast segments. For each $i \in \mathcal{V}$, we compute ordinal agreement, penalize opposite nonzero directions, and aggregate:

$$b_i = 1 - \frac{\left| s_{\text{text}}^{(i)} - s_{\text{val}}^{(i)} \right|}{6}, \qquad c_i = \begin{cases} \alpha b_i, & s_{\text{text}}^{(i)} s_{\text{val}}^{(i)} < 0, \\ b_i, & \text{otherwise}, \end{cases}$$

$$C_{\text{mag}} = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} c_i, \qquad C_{\text{risk}} = 1 - C_{\text{mag}}, \qquad \alpha = 0.75.$$

Six is the maximum ordinal distance; neutral signals do not trigger the additional directional penalty. If $\mathcal{V} = \emptyset$, the primary BGE scoring procedure returns no valid risk score, and the window is ex-

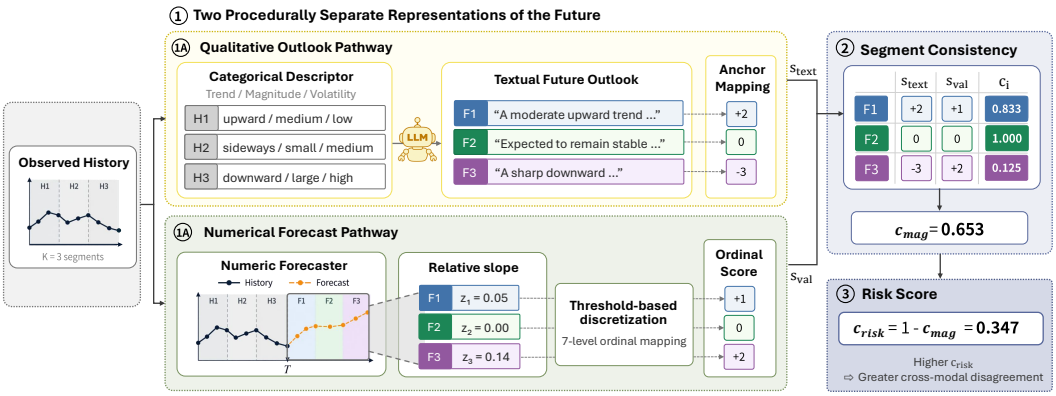


Figure 2: **Worked example of CoRE-TSF scoring.** Textual and numerical ordinal signals yield C_{risk} through segment-wise comparison ($K = 3$). All values are illustrative.

cluded from the corresponding evaluation. The score compares segment-level direction and magnitude rather than directly measuring level bias, boundary jumps, phase errors, or fine-scale oscillatory mismatches.

Single-agent and ensemble scores. For history window w and fixed forecaster b, let $C_{\text{risk}}^{(a,b)}(w)$ denote the score using textual Agent A instance a and numerical forecaster b (*Single-agent CoRE*). We combine assessments of the same forecast as *CoRE-Ensemble*:

$$C_{\text{risk}}^{(\text{ens},b)}(w) = \frac{1}{M}\sum_{a=1}^{M} C_{\text{risk}}^{(a,b)}(w), \qquad M = 4.$$

All four agent-specific scores must be available; missing BGE scores are not replaced by keyword scores. Each textual path uses the same descriptor and prompt protocol without receiving the numerical forecast. Equal weights are fixed independently of observed forecast errors. We average risk scores before ranking windows, not forecasts, agent-specific rankings, or evaluation metrics. CoRE-Ensemble is the primary method; Single-agent CoRE assesses sensitivity to the choice of Agent A. Four agents are a design choice; neither optimality nor consistent ensemble gains are claimed.

Use for selective forecasting. Both scores are ordering signals in $[0, 1]$, not calibrated probabilities of forecast failure. Within each fixed forecaster and forecasting configuration, lower-risk forecasts are retained first as coverage decreases. Ranking, tie handling, and risk–coverage and fixed-coverage evaluation are defined in Section 4.2. CoRE-TSF neither modifies Agent B's forecast nor selects among forecasting models; retention or deferral is a downstream use of the score. At scoring time, neither realized future targets nor historical forecast-error labels are required.

4 Experiments

4.1 Experimental Setup

Evaluation cohorts. We evaluate whether CoRE risk tracks forecast error (H1) and whether this ordering supports selective forecasting (H2). The primary full-horizon cohort contains 1,778 windows from 81 configurations across 22 dataset families. A configuration denotes a dataset–forecast-horizon pair, and all windows use a context length of 512. We evaluate Chronos, TimeMoE, TimesFM, Moirai, and TiRex, yielding 8,890 backbone–window instances across 405 configuration–backbone groups.

A separate matched cohort contains 891 windows from 39 configurations across the same families, restricted to $H \leq 60$. This cohort supports comparisons of the five primary backbones with LLMTime and LSTPrompt on identical windows, reflecting the practical feasibility of LLM-based forecasting in our environment. We keep matched and full-horizon results separate because their horizon support and window sets differ. Within each comparison, all reliability methods assess the

same stored point forecast for each backbone–window instance. Its realized error is fixed across methods, so differences in selective performance arise from the risk ordering rather than changes to the underlying forecasts.

Agent A models and CoRE variants. We use four Agent A models: Gemini 3.6 Flash, GPT-6 Astra, Fable 5.1 Max, and DeepSeek V4 Pro. All share the same frozen prompt, categorical history representation, segment partition, parser, and BGE-anchor mapping. CoRE-Ensemble is our primary variant, following the scoring procedure in Section 3.4 with $\alpha = 0.75$ and equal ensemble weights fixed independently of evaluation targets. Single-agent CoRE results assess sensitivity to the choice of Agent A. Implementation and model provenance appear in Appendix B, with anchor sensitivity in Appendix F.3.

Comparison methods. We compare CoRE with four baselines that estimate forecast reliability using observed history, numerical forecasts, or both, without requiring historical forecast residuals. *HistWidth* measures the width of a history-derived interval, while *PIDEV* measures forecast deviations from that range. *History-Trend* replaces the textual future outlook with historical trend and compares it with forecast-derived ordinal signals in CoRE's shared ordinal space. This control is motivated by historical–forecast consistency in forecast plausibility assessment (Nagraj et al., 2025).

Pairwise-Point measures the mean history-scale-normalized absolute distance between the target forecast and forecasts from the four other primary backbones. For the two LLM-based forecasters, LLMTime and LSTPrompt, all five primary TSFM backbones are used as auxiliaries. It uses numerical forecast disagreement as a reliability signal, motivated by ensemble-based uncertainty estimation (Lakshminarayanan et al., 2017). Our implementation uses target-centered absolute distance rather than predictive variance.

Together, these baselines test whether historical variability, history–forecast deviation, past-trend consistency, or numerical forecast disagreement can account for the reliability signal provided by CoRE's cross-modal comparison. These are study-specific controls motivated by related methodological principles rather than direct reproductions of the cited methods. Their inputs are summarized in Table 1. Random and an error-based Oracle provide reference orderings. Exact baseline definitions and their relationships to prior work are provided in Appendix D.

4.2 Evaluation Protocol

Error and rank association. Evaluation is performed separately within each configuration c and backbone b. For its $n_{c,b}$ eligible windows, define

$$e_{c,b,j} = \frac{1}{H_c} \sum_{h=1}^{H_c} |y_{c,j,h} - \hat{y}_{c,b,j,h}|, \qquad \bar{e}_{c,b} = \frac{1}{n_{c,b}} \sum_{j=1}^{n_{c,b}} e_{c,b,j}.$$

We measure the Spearman correlation $\rho_{c,b}$ between assigned risk and window-level MAE, using average ranks for ties. Positive correlation means that higher predicted risk accompanies greater realized error. Computing association within configurations avoids interpreting between-dataset differences in error magnitude as instance-level ranking ability.

Normalized selective utility. Windows are retained in increasing risk order. Let $R(q)$ denote retained MAE divided by full-coverage MAE at an attainable coverage q. At target coverage q, percentage improvement is $I(q) = 100[1 - R(\widetilde{q})]$, where $\widetilde{q} = \lceil qn \rceil / n$ is the actual retained fraction. At every attainable retained count, retained MAE is evaluated as the expectation under uniform ordering within risk ties. This rule applies to both fixed-coverage summaries and utility. Details are provided in Appendix C.

To summarize performance beyond selected coverage, we integrate R by the trapezoidal rule over all attainable coverages $1/n, \ldots, 1$, without adding a point at zero. Normalized selective utility is

$$U = \frac{A^{\mathrm{rand}} - A^{\mathrm{risk}}}{A^{\mathrm{rand}} - A^{\mathrm{oracle}}}, \qquad A^{\mathrm{rand}} = \frac{n-1}{n}.$$

Random uses the exact expected curve $R = 1$, while Oracle ranks windows by increasing realized error. Accordingly, $U = 0$ denotes random-level utility, $U = 1$ oracle-level utility, and $U < 0$ an

Table 1: Comparison of reliability signals on the primary cohort. Intervals are 95% family-bootstrap confidence intervals for U_{macro}.

Method	Signal source	ρ_{macro}	U_{macro}	95% CI for U
HistWidth	History	0.074	0.111	$[-0.024, 0.234]$
PIDEV	History + point forecast	0.016	0.016	$[-0.090, 0.108]$
History-Trend† (Nagraj et al., 2025)	History trend + point forecast	0.074	0.087	$[0.028, 0.153]$
Pairwise-Point† (Lakshminarayanan et al., 2017)	Point forecast + auxiliary forecasts	0.213	0.303	$[0.151, 0.444]$
CoRE-Ensemble	**History-derived text + point forecast**	**0.396**	**0.420**	$[0.269, 0.567]$

†Citations indicate related methodological principles; these controls are not reproductions of the cited methods. Implementation details are provided in Appendix D.

ordering worse than random, provided the normalization denominator is positive. Random's $\rho = 0$ is the expectation for an independent random ranking, not the correlation of a constant score. Utility uses the complete attainable coverage range even when a figure displays only part of that range.

Aggregation, uncertainty, and eligibility. We report macro-averages with equal weighting at each level: configurations within each backbone and family, backbones within each family, and finally the 22 families. The full aggregation equations are provided in Appendix C.

We construct percentile 95% confidence intervals using 10,000 bootstrap resamples of family-level values, preserving the multiplicity of repeated families. Method comparisons bootstrap paired family-level differences on the same eligible instances. Undefined correlations are excluded rather than replaced with zero; numerical guards and metric-specific valid counts appear in Appendix C.

4.3 H1: C_{risk} AS AN INSTANCE-LEVEL RELIABILITY SIGNAL

H1 tests whether risk correlates with forecast error within configuration–backbone groups.

Ensemble evidence. On the primary full-horizon cohort, CoRE-Ensemble achieves $\rho_{\text{macro}} = 0.396$, with a 95% family-bootstrap confidence interval of $[0.266, 0.521]$. The positive association supports the intended direction of the score: higher cross-modal disagreement tends to accompany larger realized forecast error.

The association also remains positive in the family-macro summary for each of the five forecasting backbones, ranging from 0.370 for TiRex to 0.430 for TimeMoE. Per-backbone ensemble results are provided in Appendix E.1.

Single-agent evidence. Single-agent results assess whether the observed association depends on a particular Agent A. All 20 combinations of four agents and five primary backbones have positive macro-correlation point estimates, ranging from 0.289 to 0.410.

On the matched short-horizon cohort, the eight combinations involving LLMTime and LSTPrompt have positive macro-correlation point estimates, ranging from 0.270 to 0.338. Their magnitudes should not be compared directly with full-horizon results because the evaluated windows and horizon support differ. Single-agent results for both cohorts are reported in Appendix E.2.

Comparison with alternative reliability signals. Table 1 compares CoRE with the four controls on the same 8,890 primary backbone–window instances. CoRE-Ensemble has the highest observed risk–error correlation among the listed methods, with $\rho_{\text{macro}} = 0.396$. Pairwise-Point is the strongest control on this metric, with $\rho_{\text{macro}} = 0.213$. The table also reports normalized selective utility; its interpretation and paired comparisons are discussed in Section 4.4.

Illustrative cases. Figure 3 presents selected single-agent examples illustrating the relationship between cross-modal consistency and realized forecast error. Within each selected pair, the lower-risk forecast has a smaller MAE than its higher-risk counterpart. These examples illustrate the intended interpretation of the score; they do not establish how frequently this pattern occurs across the evaluation cohort. Additional examples are provided in Appendix G.1.

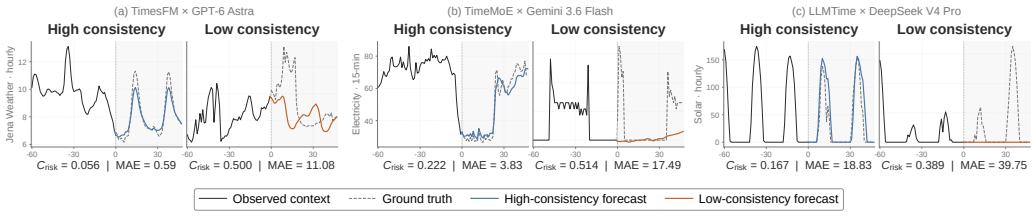


Figure 3: **Cross-modal consistency and forecast error.** High- and low-consistency single-agent examples paired within each forecaster–Agent A combination, with risk C_{risk} and realized MAE.

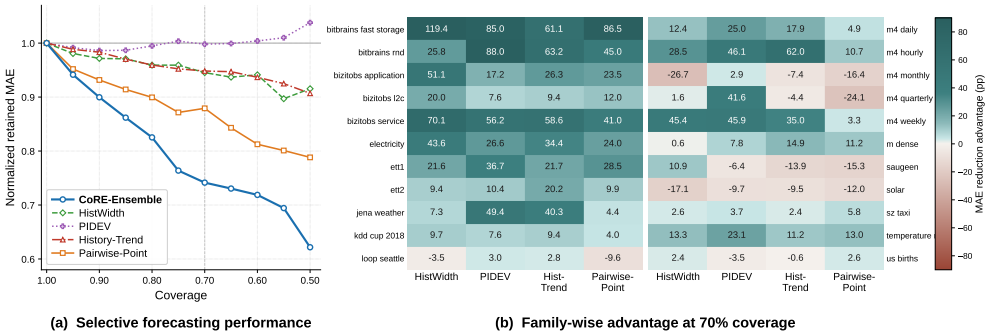


Figure 4: **(a)** Normalized retained MAE versus target coverage (lower is better). **(b)** Family-wise MAE-reduction differences between CoRE-Ensemble and each label-free comparator at 70% coverage (percentage points; positive favors CoRE-Ensemble).

4.4 H2: Risk-Ranked Selective Forecasting

All primary comparisons use the same 8,890 backbone–window instances and the evaluation protocol in Section 4.2.

Aggregate selective utility. As shown in Table 1, CoRE-Ensemble achieves $U_{\text{macro}} = 0.420$ (95% CI $[0.269, 0.567]$), indicating positive selective utility relative to random retention. Pairwise-Point has the highest utility among the four controls, with $U_{\text{macro}} = 0.303$. CoRE's paired utility difference from History-Trend is $+0.333$ (95% CI $[0.168, 0.491]$), supporting a benefit over this past-trend consistency control. Its difference from Pairwise-Point is $+0.117$, but the paired 95% CI $[-0.104, 0.338]$ includes zero. Thus, CoRE has a higher observed mean utility without a statistically resolved advantage over Pairwise-Point.

Coverage dependence and family-level effects. Figure 4(a) compares risk–coverage curves on the same cohort as the baseline table. The vertical axis is hierarchically averaged relative retained MAE, with lower values indicating better selective performance. Random has the exact expected value $R = 1$, and all methods coincide at full coverage.

Across the evaluated non-full coverage levels shown in the figure, CoRE has lower aggregate relative retained MAE than the four controls, with Pairwise-Point providing the closest comparison. History-Trend and HistWidth yield similar fixed-coverage performance: their improvements at 70% coverage are 5.2% and 5.5%, respectively, despite their different full-curve utilities. PIDEV exceeds $R = 1$ at several displayed coverage levels, indicating that its ordering can increase retained error.

Panel (b) presents a split heatmap of family-level improvements relative to each of the four controls. For family f and baseline m, each cell reports

$$D_{f,m} = I_f^{\text{CoRE}}(0.7) - I_f^{m}(0.7), \quad m \in \{\text{HistWidth}, \text{PIDEV}, \text{History-Trend}, \text{Pairwise-Point}\}.$$

Differences are expressed in percentage points. Positive values favor CoRE, while negative values favor the corresponding baseline. All 22 families and the same four comparators are shown, without selecting a different comparator for each family.

Table 2: MAE of CoRE-Ensemble (CoRE) and Pairwise-Point (PW) at selected coverages. All models and methods share 24 windows per dataset. Horizons: 48 steps for ETTh1, ETTh2, and Jena Weather; 13 for M4-Weekly. Parentheses show percentage improvement over full-coverage MAE.

Source	Model	Method	MAE (Improvement, %) Coverage 100%	85%	70%
ETTh1/H	Chronos	**CoRE**	6.10(0.0%)	**4.51(26.1%)**	**3.34(45.3%)**
		PW	6.10(0.0%)	4.91(19.5%)	4.99(18.3%)
	TimeMoE	**CoRE**	5.73(0.0%)	**4.49(21.7%)**	**2.83(50.7%)**
		PW	5.73(0.0%)	6.17(−7.6%)	6.72(−17.2%)
	LSTPrompt	**CoRE**	11.23(0.0%)	**9.29(17.2%)**	**6.21(44.7%)**
		PW	11.23(0.0%)	11.10(1.1%)	9.83(12.4%)
ETTh2/H	Chronos	**CoRE**	6.22(0.0%)	**5.96(4.1%)**	**5.49(11.7%)**
		PW	6.22(0.0%)	6.44(−3.6%)	6.63(−6.6%)
	TimeMoE	**CoRE**	6.28(0.0%)	**5.89(6.3%)**	**5.46(13.0%)**
		PW	6.28(0.0%)	6.48(−3.2%)	6.40(−2.0%)
	LSTPrompt	**CoRE**	9.07(0.0%)	8.58(5.4%)	**6.76(25.5%)**
		PW	9.07(0.0%)	**8.15(10.1%)**	7.92(12.7%)

Source	Model	Method	MAE (Improvement, %) Coverage 100%	85%	70%
Jena Weather/H	Chronos	**CoRE**	6.42(0.0%)	**2.99(53.4%)**	**1.89(70.6%)**
		PW	6.42(0.0%)	3.44(46.4%)	3.68(42.7%)
	TimeMoE	**CoRE**	6.71(0.0%)	4.22(37.0%)	**3.49(48.0%)**
		PW	6.71(0.0%)	**3.07(54.2%)**	3.59(46.5%)
	LSTPrompt	**CoRE**	6.22(0.0%)	**4.00(35.7%)**	**2.60(58.2%)**
		PW	6.22(0.0%)	4.34(30.3%)	4.85(22.0%)
M4-Weekly/W	Chronos	**CoRE**	206.68(0.0%)	**139.43(32.5%)**	112.43(45.6%)
		PW	206.68(0.0%)	198.64(3.9%)	**79.24(61.7%)**
	TimeMoE	**CoRE**	657.92(0.0%)	**213.70(67.5%)**	**190.69(71.0%)**
		PW	657.92(0.0%)	229.90(65.1%)	204.61(68.9%)
	LSTPrompt	**CoRE**	264.30(0.0%)	**204.26(22.7%)**	**169.61(35.8%)**
		PW	264.30(0.0%)	264.19(0.0%)	268.51(−1.6%)

Against Pairwise-Point, CoRE has a positive improvement difference in 17 families and a negative difference in 5. The mean difference is approximately $+11.5$ percentage points. These results show the aggregate advantage coexists with substantial variation in relative performance across families.

Absolute MAE at selected coverages. Table 2 reports absolute forecast error on four selected datasets: ETTh1, ETTh2, Jena Weather, and M4-Weekly. We compare CoRE-Ensemble and Pairwise-Point at 100%, 85%, and 70% target coverage using two of the five time-series foundation models (Chronos and TimeMoE) and one LLM-based forecaster (LSTPrompt). Parentheses report percentage improvement over each model's full-coverage MAE.

CoRE achieves lower MAE than Pairwise-Point in 21 of the 24 reduced-coverage comparisons shown. The largest relative reduction is 57.9%, observed on ETTh1 with TimeMoE at 70% target coverage. These selected results illustrate how risk-based retention translates into changes in absolute forecast error. Complete comparison tables across all 22 dataset families and evaluated models are provided in the appendix.

5 Conclusion

We proposed CoRE-TSF, a framework for assessing the reliability of individual time-series forecasts through consistency between an LLM's qualitative outlook and a numerical forecast. CoRE-TSF assigns output-level risk scores without modifying forecasts or requiring model internals or historical forecast-error labels. Across the evaluated forecasters and Agent A models, we observed a positive association between cross-modal disagreement and forecast error on average, and showed retaining lower-risk forecasts reduced retained error relative to random retention. CoRE-Ensemble achieved higher observed mean selective utility than Pairwise-Point and lower retained error in 17 of 22 dataset families at 70% coverage. Although comparative benefits varied across families and coverage levels, these findings support cross-modal consistency as a signal for guiding forecast retention and deferral. This work contributes an output-level reliability assessment method for existing forecasters, with empirical evidence for using textual–numerical consistency to assess forecast reliability.

Limitations. CoRE-TSF's segment-level direction–magnitude comparison does not directly capture level bias, phase errors, or fine-scale mismatches; scores may depend on categorical history descriptors and ordinal mapping. Both LLM-based forecasters were evaluated only at short horizons. Results therefore support empirical reliability ranking under evaluated conditions. Future work could capture broader forecast-error types, evaluate longer horizons, and reduce LLM inference costs.

AI USE STATEMENT

In this work, generative AI tools were used to provide feedback on research methodology and experimental design, assist with the implementation and debugging of analysis code, and support the interpretation and consistency checking of experimental results. They were also used to improve the clarity and organization of the manuscript, edit LaTeX source, and assist with literature summarization.

All research questions, experimental decisions, evaluation protocols, statistical analyses, and scientific claims were determined and verified by the authors. AI-assisted code was reviewed and tested by the authors, all reported numerical results were checked against the underlying experimental outputs, and cited sources were verified against the original publications. The authors take responsibility for the final content of this work, including any text, code, or artifacts produced with the assistance of generative AI.

REFERENCES

Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. Gift-eval: A benchmark for general time series forecasting model evaluation. *arXiv preprint arXiv:2410.10393*, 2024.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

Andreas Auer, Patrick Podest, Daniel Klotz, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. Tirex: Zero-shot forecasting across long and short horizons with enhanced in-context learning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 57529–57580. Curran Associates, Inc., 2025. doi: 10.52202/085713-1929. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/5356603f9c47399adfd372f77a677057-Paper-Conference.pdf.

George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung. *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, Hoboken, NJ, 5 edition, 2015.

Jonas Brusokas, Seshu Tirupathi, Dalin Zhang, and Torben Bach Pedersen. The time-energy model: selective time-series forecasting using energy-based models. *Transactions on Machine Learning Research*, 2025.

Jonas Brusokas, Seshu Tirupathi, and Torben Bach Pedersen. Fitem: Fine-tuning time-series foundation models for selective forecasting. In Vincent Lemaire, Georgiana Ifrim, Anthony Bagnall, Simon Malinowski, Patrick Schäfer, and Romain Tavenard, editors, *Advanced Analytics and Learning on Temporal Data*, pages 181–195, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-15535-1.

Peng Chen, Yihang Wang, Yang Shu, Yunyao Cheng, Kai Zhao, Zhongwen Rao, Lujia Pan, Bin Yang, and Chenjuan Guo. Cc-time: Cross-model and cross-modality time series forecasting. *arXiv preprint arXiv:2508.12235*, 2025.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 10148–10167. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/das24c.html.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024. doi: 10.1038/s41586-024-07421-0. URL https://doi.org/10.1038/s41586-024-07421-0.

Ninghui Feng, Songning Lai, Xin Zou, Jiayu Yang, Kunlong Feng, Fobao Zhou, Yutao Yue, Zheng Chu, Xinan Chen, David Chieng, and Hang Zhao. Towards reliable time series forecasting under future uncertainty: Ambiguity and novelty rejection mechanisms. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5126–5130, 2026. doi: 10.1109/ICASSP55912.2026.11464741.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *International conference on machine learning*, pages 2151–2159. PMLR, 2019.

Isaac Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021.

Tilmann Gneiting and Matthias Katzfuss. Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1(1):125–151, 2014. doi: 10.1146/annurev-statistics-062713-085831. URL https://doi.org/10.1146/annurev-statistics-062713-085831.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(2): 243–268, 2007. doi: 10.1111/j.1467-9868.2007.00587.x.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. Large language models are zero-shot time series forecasters. *Advances in neural information processing systems*, 36:19622–19635, 2023.

Rob J. Hyndman, Anne B. Koehler, J. Keith Ord, and Ralph D. Snyder. *Forecasting with Exponential Smoothing: The State Space Approach*. Springer Series in Statistics. Springer, Berlin, Heidelberg, 2008. doi: 10.1007/978-3-540-71918-2.

Ricardo Inácio, Vitor Cerqueira, Marília Barandas, and Carlos Soares. Selective time series forecasting via metalearning, 2026. URL https://arxiv.org/abs/2606.23448.

Furong Jia, Kevin Wang, Yixiang Zheng, Defu Cao, and Yan Liu. Gpt4mts: Prompt-based large language model for multimodal time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 23343–23351, 2024.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728*, 2023.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Bryan Lim, Sercan O. Arik, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting, 2020. URL https://arxiv.org/abs/1912.09363.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words, 2022. URL https://arxiv.org/abs/2205.14334.

Chenxi Liu, Qianxiong Xu, Hao Miao, Sun Yang, Lingzheng Zhang, Cheng Long, Ziyue Li, and Rui Zhao. TimeCMA: Towards LLM-empowered multivariate time series forecasting via cross-modality alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 18780–18788, 2025a.

Haoxin Liu, Zhiyuan Zhao, Jindong Wang, Harshavardhan Kamarthi, and B Aditya Prakash. Lst-prompt: Large language models as zero-shot time series forecasters by long-short-term prompting. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7832–7840, 2024.

Haoxin Liu, Zhiyuan Zhao, Shiduo Li, and B. Aditya Prakash. Evaluating system 1 vs. 2 reasoning approaches for zero-shot time series forecasting: A benchmark and insights, 2025b. URL https://arxiv.org/abs/2503.01895.

Peiyuan Liu, Hang Guo, Tao Dai, Naiqi Li, Jigang Bao, Xudong Ren, Yong Jiang, and Shu-Tao Xia. CALF: Aligning LLMs for time series forecasting via cross-modal fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 18915–18923, 2025c.

Zhiding Liu, Mingyue Cheng, Guanhao Zhao, Jiqian Yang, Qi Liu, and Enhong Chen. Improving time series forecasting via instance-aware post-hoc revision. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 35604–35629. Curran Associates, Inc., 2025d. doi: 10.52202/085713-1196. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/331c41353b053683e17f7c88a797701d-Paper-Conference.pdf.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

V.P. Nagraj, Amy E. Benefield, Desiree Williams, and Stephen D. Turner. PLANES: Plausibility analysis of epidemiological signals. *PLOS ONE*, 20(3):e0320442, 03 2025. doi: 10.1371/journal.pone.0320442. URL https://doi.org/10.1371/journal.pone.0320442.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers, 2023. URL https://arxiv.org/abs/2211.14730.

Boris N. Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-beats: Neural basis expansion analysis for interpretable time series forecasting, 2020. URL https://arxiv.org/abs/1905.10437.

Junwoo Park, Hyuck Lee, Dohyun Lee, Daehoon Gwak, and Jaegul Choo. Revisiting llms as zero-shot time series forecasters: Small noise can break large models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 906–922, 2025.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

David Salinas, Valentin Flunkert, and Jan Gasthaus. Deepar: Probabilistic forecasting with autoregressive recurrent networks, 2019. URL https://arxiv.org/abs/1704.04110.

Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-moe: Billion-scale time series foundation models with mixture of experts. In *International conference on learning representations*, volume 2025, pages 34635–34667, 2025.

Tejas Srinivasan, Jack Hessel, Tanmay Gupta, Bill Yuchen Lin, Yejin Choi, Jesse Thomason, and Khyathi Chandu. Selective "selective prediction": Reducing unnecessary abstention in vision-language reasoning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12935–12948, Bangkok, Thailand,

August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.767. URL https://aclanthology.org/2024.findings-acl.767/.

Kamile Stankeviciute, Ahmed M. Alaa, and Mihaela van der Schaar. Conformal time-series forecasting. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 6216–6228. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/312f1ba2a72318edaaa995a67835fad5-Paper.pdf.

Luca Stradiotti, Laurens Devos, Anna Monreale, Jesse Davis, and Andrea Pugnana. Bounded-abstention multi-horizon time-series forecasting, 2026. URL https://arxiv.org/abs/2602.04714.

Guangzhi Sun, Potsawee Manakul, Adian Liusie, Kunat Pipatanakul, Chao Zhang, Phil Woodland, and Mark Gales. Crosscheckgpt: Universal hallucination ranking for multimodal foundation models, 2024. URL https://arxiv.org/abs/2405.13684.

Mingtian Tan, Mike A Merrill, Vinayak Gupta, Tim Althoff, and Thomas Hartvigsen. Are language models actually useful for time series forecasting? *Advances in Neural Information Processing Systems*, 37:60162–60191, 2024.

Xiaoyu Tao, Mingyue Cheng, Ze Guo, Shuo Yu, Yaguo Liu, Qi Liu, and Shijin Wang. Memcast: Memory-driven time series forecasting with experience-conditioned reasoning, 2026. URL https://arxiv.org/abs/2602.03164.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.330. URL https://aclanthology.org/2023.emnlp-main.330/.

Shivani Tomar, Seshu Tirupathi, Elizabeth Daly, and Ivana Dusparic. Shapelets-enriched selective forecasting using time series foundation models, 2026. URL https://arxiv.org/abs/2601.11821.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, NY, 2005. doi: 10.1007/b106715.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023. URL https://arxiv.org/abs/2203.11171.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. *arXiv preprint arXiv:2402.02592*, 2024.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-pack: Packed resources for general chinese embeddings, 2024. URL https://arxiv.org/abs/2309.07597.

Chen Xu and Yao Xie. Conformal prediction for time series. *IEEE transactions on pattern analysis and machine intelligence*, 45(10):11575–11587, 2023.

Hao Xue and Flora D Salim. Promptcast: A new prompt-based learning paradigm for time series forecasting. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):6851–6864, 2023.

Jiaxin Zhang, Zhuohang Li, Kamalika Das, Bradley Malin, and Sricharan Kumar. SAC^3: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15445–15458, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.1032. URL https://aclanthology.org/2023.findings-emnlp.1032/.

Xiang Zhang, Senyu Li, Ning Shi, Bradley Hauer, Zijun Wu, Grzegorz Kondrak, Muhammad Abdul-Mageed, and Laks V. S. Lakshmanan. Cross-modal consistency in multimodal large language models, 2024. URL https://arxiv.org/abs/2411.09273.

Shiqiao Zhou, Holger Schöner, Huanbo Lyu, Edouard Fouché, and Shuo Wang. Balm-tsf: Balanced multimodal alignment for llm-based time series forecasting. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 4498–4508, 2025.

A Evaluation Cohorts and Dataset Details

A.1 Benchmark Source and Definitions

All time-series windows are drawn from GIFT-Eval (Aksu et al., 2024), a curated collection of public forecasting benchmarks spanning diverse domains, frequencies, and horizon lengths. We use the term *configuration* to mean a unique (dataset, H) pair, where H is the forecast horizon. Forecasting backbones and Agent A variants are *experimental factors*, not dimensions of a configuration. The *dataset family* (e.g., ETT, Weather, M3-Monthly) is the highest-level macro-aggregation unit; families pool the configurations that share a common source and statistical character. Every window uses a context length of $T = 512$ steps. All backbones within the same cohort operate on exactly the same set of windows (the *common-window* principle), so cross-backbone comparisons do not confound window selection.

A.2 Primary Full-Horizon Cohort

The *primary cohort* consists of **1,778 windows** drawn from **81 configurations** spanning **22 dataset families**. No restriction is placed on the forecast horizon H. Five zero-shot time-series forecasting models serve as Agent B backbones in this cohort: Chronos, TimeMoE, TimesFM, Moirai, and TiRex. The manifest is frozen at a single version and shared identically across all five backbones and all Agent A variants; no backbone-specific re-sampling was performed after the manifest was finalised.

A.3 Matched $H \leq 60$ Cohort

The *matched cohort* is a subset of the primary manifest restricted to windows where $H \leq 60$, yielding **891 windows** from **39 configurations** across **22 families**. The 39 configurations are the subset of the 81 primary configurations for which at least one window has $H \leq 60$; the remaining 42 configurations have exclusively $H > 60$ windows and are absent from the matched cohort. Two autoregressive LLM-based Agent B backbones (LLMTime and LSTPrompt) operate exclusively on this cohort; we use the matched $H \leq 60$ cohort for feasible and comparable evaluation of both backbones. Because the two cohorts differ in both horizon distribution and configuration support, their metric values should not be directly compared as measures of backbone differences alone.

A.4 Sampling, Inclusion, and Exclusion Accounting

Windows were sampled from each eligible configuration with a maximum cap of **24 windows per configuration**, using a chronologically ordered cutoff sequence. The frozen manifest contains 1,778 windows across 81 configurations and 22 dataset families, all of which are feasible for the five primary backbones. Context length is $T = 512$ for all windows.

The eligibility audit covers **97 candidate configurations**: 66 are fully eligible, 15 are partially eligible (fewer than 24 eligible windows; included with all available windows), and 16 are ineligible. Ineligible configurations were excluded for one of two reasons: no eligible window (15 configurations, because the target length was shorter than the prediction horizon or no series had sufficient history at any evaluation cutoff) or an entirely missing context region (one hierarchical sales configuration). Cutoff dates follow the official GIFT-Eval evaluation protocol; spacing between cutoffs is frequency-dependent and not uniform across configurations. Context and target are taken directly from the GIFT-Eval data fields: `history_value` (last $T = 512$ points) for context and `future_value` (H steps) for the forecast target.

The following tables summarise the final cohort composition. Table 3 gives overall counts per cohort, and Table 4 gives the per-family breakdown.

Table 3: Cohort summary.

Cohort	Windows	Configs	Families	T	Horizon	Backbones
Primary	1,778	81	22	512	unrestricted	Chronos, TimeMoE, TimesFM, Moirai, TiRex
Matched ($H \leq 60$)	891	39	22	512	$H \leq 60$	all five above[a] + LLMTime, LSTPrompt

[a]The five primary TSFM backbones (Chronos, TimeMoE, TimesFM, Moirai, TiRex) are also evaluated on the matched cohort as a subset of the primary manifest. LLMTime and LSTPrompt are the two LLM-based backbones exclusive to the matched cohort.

Table 4: Dataset-family summary: frequency, horizon range, configurations and windows per cohort. "P-cfgs" = primary-cohort configurations (full-horizon); "M-cfgs" = matched-cohort configurations ($H \leq 60$); "P-wins" and "M-wins" are the corresponding window counts. $T = 512$ for all windows. †Families with configurations at multiple frequencies; the listed frequency is the most common.

Family	Freq.	H range	P-cfgs	M-cfgs	P-wins	M-wins	Domain
bitbrains_fast_storage†	5T	48–720	4	2	96	48	Web/Cloud
bitbrains_rnd†	5T	48–720	4	2	96	48	Web/Cloud
bizitobs_application	10S	60–900	3	1	30	24	Web/Cloud
bizitobs_l2c†	5T	48–720	6	2	110	48	Web/Cloud
bizitobs_service	10S	60–900	3	1	72	24	Web/Cloud
electricity†	15T	30–720	7	3	168	72	Energy
ett1†	15T	30–720	7	3	162	69	Energy
ett2†	15T	30–720	7	3	162	69	Energy
jena_weather†	10T	48–720	6	2	144	48	Nature
kdd_cup_2018	H	48–720	3	1	33	24	Nature
loop_seattle†	5T	48–720	6	2	144	48	Transport
m4_daily	D	14	1	1	24	24	Econ/Fin
m4_hourly	H	48	1	1	24	24	Econ/Fin
m4_monthly	M	18	1	1	24	24	Econ/Fin
m4_quarterly	Q	8	1	1	24	24	Econ/Fin
m4_weekly	W	13	1	1	24	24	Econ/Fin
m_dense†	H	30–720	4	2	96	48	Transport
saugeen†	D	8–30	3	3	47	47	Nature
solar†	10T	48–720	6	2	144	48	Energy
sz_taxi†	15T	48–720	4	2	96	48	Transport
temperature_rain	D	30	1	1	24	24	Nature
us_births†	D	8–30	2	2	34	34	Healthcare
Total			**81**	**39**	**1,778**	**891**	

B Implementation Details

B.1 Categorical History Representation and Prompt

The same frozen input representation and prompt are used in all main experiments. The context window is encoded as a sequence of $K = 3$ qualitative segment descriptors, which constitute the sole input to Agent A. Agent A never receives the raw numerical values or Agent B's numerical forecast. The prompt does not include the forecast horizon, sampling frequency, timestamps, or future-segment lengths. Future segments are specified only by their order and matched to numerical segments by index.

Each descriptor corresponds to one segment of the observed context and contains three categorical attributes. *Direction* is determined by the sign and magnitude of the first-order least-squares slope over the segment:

$$\text{direction} = \begin{cases} \texttt{upward} & \text{if slope} > 0.02, \\ \texttt{downward} & \text{if slope} < -0.02, \\ \texttt{sideways} & \text{otherwise.} \end{cases}$$

Magnitude is determined from the absolute slope:

$$\text{magnitude} = \begin{cases} \texttt{small} & \text{if } |\text{slope}| < 0.05, \\ \texttt{medium} & \text{if } 0.05 \le |\text{slope}| < 0.15, \\ \texttt{large} & \text{otherwise.} \end{cases}$$

Volatility is determined from the within-segment standard deviation:

$$\text{volatility} = \begin{cases} \texttt{low} & \text{if } \sigma < 0.5, \\ \texttt{medium} & \text{if } 0.5 \le \sigma < 2.0, \\ \texttt{high} & \text{otherwise,} \end{cases}$$

where σ is the population standard deviation, computed by dividing the sum of squared deviations by the segment length. The slope and standard deviation in these formulae are computed from the raw observed values of each context segment, without prior normalization; the thresholds therefore apply to values in the original series units. These thresholds apply only to Agent A's context-side feature extraction and are separate from the forecast-slope thresholds used to construct s_{val} in Section B.6.

Deterministic equal-split convention. The main setting is $K = 3$. We apply the same deterministic partitioning convention to the observed context and the numerical forecast horizon. For a sequence of length L, where $L = T$ for the context and $L = H$ for the forecast, write

$$L = aK + r, \qquad 0 \le r < K.$$

The first r segments contain $a+1$ consecutive time steps, and the remaining $K-r$ segments contain a time steps. The resulting segments are disjoint, preserve chronological order, and jointly cover the entire sequence. When $L < K$, any empty numerical segments are treated as degenerate under the convention defined in Section 3.4.

For Agent A, the K context descriptors are serialized in chronological order and concatenated into a single structured prompt. The full prompt template for $K = 3$ is reproduced verbatim below.

```
Task: Write forecast reasoning for future segments.

Rules:
- Use exactly one sentence per segment.
- Do not explain the task.
- Do not give definitions.
- Do not generate numbers.
- Use only qualitative trend language.
- Output only the final segment lines.
- Output exactly 3 segment lines.
- Stop after Segment 3.
- Write the sentences based only on the structural description below.

Output format:
Segment 1: <one sentence>
Segment 2: <one sentence>
Segment 3: <one sentence>

Structural description:

Segment 1:
- Direction: <direction>
- Magnitude: <magnitude>
- Volatility: <volatility>

Segment 2:
- Direction: <direction>
- Magnitude: <magnitude>
- Volatility: <volatility>

Segment 3:
- Direction: <direction>
- Magnitude: <magnitude>
- Volatility: <volatility>

Now write the output, starting with "Segment 1:".
```

B.2 Agent A Inference and Output Parsing

The four Agent A models used in the main experiments are listed in Table 5 in their primary reporting order. All four models use the frozen prompt (Section B.1) with $K = 3$. Stored responses cover all 1,778 primary-cohort windows for each Agent A model. Responses for the 891 matched-cohort windows are reused from this set.

Table 5: Agent A model roster. All four models use the frozen prompt with $K = 3$ and process all 1,778 primary-cohort windows; matched-cohort consistency scores reuse the same stored outputs for the $H \leq 60$ subset (891 of the 1,778 windows).

Order	Display name	Model identifier	Provider
1	Gemini 3.6 Flash	`gemini-3.6-flash`	Google
2	GPT-6 Astra	`gpt-6-astra`	OpenAI
3	Fable 5.1 Max	`claude-fable-5-1`	Anthropic
4	DeepSeek V4 Pro	`deepseek-v4-pro`	DeepSeek

Agent A inference uses one completion per window, without multi-sample averaging. All completions are produced in a single turn without conversational history. The model-specific inference settings are as follows:

- *Gemini 3.6 Flash*: temperature 1.0, `thinking_budget` $= 1$, and `max_tokens` $= 2{,}048$.
- *DeepSeek V4 Pro*: temperature 1.0, top-$p = 0.95$, `max_completion_tokens` $= 32{,}768$, and `reasoning_effort` set to `max`.

- *GPT-6 Astra*: `max_completion_tokens` was set to 2,048; temperature setting is not recorded in the stored outputs and cannot be verified.
- *Fable 5.1 Max*: `max_tokens` was set to 16,384; temperature was not explicitly configured and is unverified from stored outputs.

No explicit random seed was supplied in the Agent A API calls. Accordingly, reproducibility of these outputs relies on the preserved raw responses rather than deterministic regeneration from a fixed sampling seed.

Agent A's output is parsed by extracting lines matching the pattern `Segment N:` ⟨text⟩ using a regular expression that is case-insensitive and spans newlines. Segments are expected in ascending order (1 through K). The parser validates that exactly K segments are present and that indices are in order; outputs failing either check are treated as invalid.

Invalid-response policy: when the parser fails to extract exactly K segments in the correct order, the output is marked invalid and excluded from all metric computation. No retry is performed for parse failures. Network and API errors trigger automatic retries before the window is marked as failed. A keyword-based fallback scorer is available for diagnostic testing when the BGE model is unavailable, but it is not used in the primary BGE-based evaluation. All four Agent A models in the main results use the BGE scorer exclusively.

B.3 BGE-Based Text-to-Signal Mapping

Agent A segment texts are mapped to ordinal scores via the **`BAAI/bge-small-en-v1.5`** sentence encoder. The mapping assigns each text to the level $\ell^* \in \{-3, -2, -1, 0, +1, +2, +3\}$ whose anchor sentences have the highest maximum cosine similarity to the query:

$$s_{\text{text}} = \arg\max_{\ell} \max_{a \in \mathcal{A}_\ell} \cos(g(\text{text}),\, g(a)),$$

where $g(\cdot)$ denotes the L2-normalised BGE embedding and $\mathcal{A}_\ell$ is the set of anchor sentences at level ℓ. In the event of a tie in maximum cosine similarity across levels, the assigned level is the one whose key appears first in the anchor dictionary iteration order ($+3, +2, +1, 0, -1, -2, -3$).

Table 6 lists the complete anchor bank. Levels $\pm 3, \pm 2, \pm 1$ each have five anchor sentences; level 0 has nineteen sentences, reflecting iterative calibration adjustments to capture stable-forecast phrasings produced by GPT-4o-mini. The same fixed anchor bank is used for all four Agent A models.

Table 6: Complete BGE anchor bank. Each entry is one anchor sentence; Agent A output is assigned to the level with the highest maximum cosine similarity across that level's entries.

Level	Anchor sentences
$+3$	(1) A strong upward trend is expected to continue with significant momentum. (2) The series is likely to surge sharply higher. (3) A steep and dramatic increase is anticipated. (4) Strong upward momentum is expected to drive values significantly higher. (5) A powerful and accelerating rise is likely to continue.
$+2$	(1) A moderate upward trend is likely to continue. (2) Continued moderate growth is expected. (3) The series is expected to rise steadily at a moderate pace. (4) A consistent upward movement at a moderate rate is anticipated. (5) Reliable upward progress at a moderate level is expected.
$+1$	(1) A tentative firming is expected, with the series making a modest upward advance. (2) A marginal increase is likely. (3) The series is likely to inch upward, posting modest gains over time. (4) The series shows a weak bullish lean, drifting marginally higher with tentative support. (5) A subdued upward movement is expected.
0	(1) The series is expected to remain stable. (2) No meaningful directional trend is anticipated. (3) The series is likely to hover around current levels. (4) A sideways movement with no clear direction is expected. (5) Stabilization around current levels is likely. (6) The market is expected to maintain a stable trajectory with no directional movement. (7) A continuation of the current state is anticipated, with no significant directional change. (8) A persistent equilibrium is expected, with no clear upward or downward bias. (9) Conditions are likely to remain flat, with minimal fluctuations and no trend. (10) The series is expected to drift sideways, maintaining a stable and directionless path. (11) A continued period of stability is expected, with no meaningful directional change. (12) Consistent performance is expected, reflecting a sideways movement with no clear trend. (13) Stability is expected to persist, characterized by no net directional movement. (14) The series is expected to remain stable with low variability and no trend in either direction. (15) A continued period of stability is expected, characterized by slight movements but no directional trend. (16) Trends suggest stability will persist, with slight variations and no upward or downward movement. (17) The environment appears poised to remain stable, with no directional bias in performance. (18) Overall conditions are likely to remain consistent, with low levels of change and no clear trend. (19) Stability is expected to continue, characterized by minor fluctuations without directional momentum.
-1	(1) A mild softening is expected, with the series easing back slightly from current levels. (2) A mild decline is likely. (3) The series is likely to recede gradually, posting modest losses over time. (4) A minor softening is anticipated. (5) The series shows a weak bearish lean, drifting marginally lower without conviction.
-2	(1) A steady deterioration is expected, with the series losing ground at a consistent moderate pace. (2) A steady decline at a moderate pace is likely. (3) The series is expected to fall at a moderate rate. (4) A clear downward trend at a measured rate is anticipated. (5) Continued moderate deterioration is expected.
-3	(1) A sharp and significant decline is expected. (2) The series is likely to drop steeply. (3) A dramatic downward movement is anticipated. (4) Strong downward pressure is expected to drive values significantly lower. (5) A severe and accelerating fall is likely to continue.

B.4 KEYWORD-BASED FALLBACK

When a BGE scorer is not available (e.g., during unit testing), a deterministic keyword scorer is available as a separate diagnostic path. The fallback checks for stable-class keywords first (e.g., "stable", "sideways", "hover"), then upward keywords, then downward keywords, and applies intensity qualifiers to distinguish levels within each direction. The primary evaluation uses BGE exclusively; a BGE failure does not trigger substitution of keyword scores. The keyword implementation assigns $C_{\text{mag}} = 0$ and $C_{\text{risk}} = 1$ to all-degenerate windows, whereas the primary BGE evaluation treats them as invalid.

B.5 FORECASTING BACKBONES AND POINT-FORECAST CONSTRUCTION

Table 7 summarises inference parameters for all seven Agent B backbones. All backbones produce a single aggregated one-dimensional point forecast that serves as the input to the numerical signal extraction pipeline (§B.6).

Table 7: Agent B backbone inference parameters. LLMTime and LSTPrompt use the matched $H \leq 60$ cohort; the five TSFMs are evaluated on the primary cohort and its matched subset.

Backbone	Cohort	Checkpoint / identifier	Samples	Point-forecast agg.
Chronos	Primary	`amazon/chronos-t5-large`	20	position-wise median
TimeMoE	Primary	`Maple728/TimeMoE-50M`	1 (deterministic)	direct output
TimesFM	Primary	`google/timesfm-1.0-200m-pytorch`	1 (deterministic)	direct output
Moirai	Primary	`Salesforce/moirai-1.1-R-base`	20	position-wise median
TiRex	Primary	`NX-AI/TiRex`	1 (model mean)	model mean
LLMTime	Matched	LLaMA-2 via llamacpp	20	position-wise median
LSTPrompt	Matched	Ollama-hosted model	5	position-wise median

For backbones that generate multiple samples (Chronos, Moirai, LLMTime, LSTPrompt), the *position-wise median* across samples is computed first to obtain a single 1-D point forecast of length H. This aggregated forecast is then passed to the slope-based numerical signal extraction pipeline.

Chronos uses bfloat16 precision and probabilistic decoding (20 samples, position-wise median). Moirai inputs are float32 and uses probabilistic decoding (20 samples, position-wise median). TimeMoE uses `torch_dtype="auto"` (bfloat16 on compatible hardware) and deterministic autoregressive decoding. TimesFM uses deterministic decoding. TiRex uses float32 context and outputs the model mean directly. LSTPrompt uses plain-text completion-mode prompting with temperature 0.8 and 5 samples; completions are position-wise median-aggregated.

Input normalization. Backbones differ in whether the inference wrapper applies explicit normalization before model input:

- *TimeMoE*: the context is z-score normalized in the inference wrapper using the context mean μ and standard deviation σ (with σ clipped to at least 10^{-8}) before being passed to the model; the output is inverse-transformed as $\hat{y} = \hat{y}_{\text{norm}} \cdot \sigma + \mu$.
- *LLMTime*: context-derived normalization is applied internally by the llmtime library with scale parameters $\alpha = 0.99$ and $\beta = 0.3$. The tokenized numerical forecast is generated as 20 samples and inverse-transformed to the original scale; the point forecast is the position-wise sample median.
- *LSTPrompt*: an LLMTime-compatible scaler is applied in the inference wrapper with the same parameters ($\alpha = 0.99$, $\beta = 0.3$, `base`$= 10$, `prec`$= 3$) before serialization and after deserialization.
- *Chronos*, *TimesFM*, *Moirai*, *TiRex*: no explicit normalization is applied in the inference wrapper; raw context values are passed directly to the backbone's own inference pipeline, which handles any internal scaling.

B.6 Numerical Signal Extraction and Scoring Conventions

The numerical trend score s_{val} for each forecast segment is computed from the aggregated one-dimensional point forecast as follows. All seven backbones use the same shared segment boundaries derived from the forecast horizon under the equal-split rule with $K = 3$ segments. For each segment i the first-order least-squares slope is computed:

$$\hat{m}_i = \text{polyfit}(x,\, \mathbf{y}_i,\, 1)[0], \quad x = [0, 1, \ldots, |\mathbf{y}_i| - 1],$$

and normalised using the segment mean:

$$z_i = \frac{\hat{m}_i}{|\bar{y}_i| + \varepsilon}, \quad \varepsilon = 10^{-8}.$$

The denominator $|\bar{y}_i|$ is the absolute mean of the *forecast* segment (not the context mean). The normalised slope is then mapped to an ordinal score via fixed thresholds:

$$s_{\text{val}}^{(i)} = \begin{cases} -3, & z_i \le -0.20, \\ -2, & -0.20 < z_i \le -0.08, \\ -1, & -0.08 < z_i < -0.02, \\ 0, & -0.02 \le z_i < 0.02, \\ +1, & 0.02 \le z_i < 0.08, \\ +2, & 0.08 \le z_i < 0.20, \\ +3, & z_i \ge 0.20. \end{cases} \tag{1}$$

At the negative thresholds, $z_i = -0.08$ maps to -2 and $z_i = -0.20$ maps to -3, whereas $z_i = -0.02$ maps to 0.

A segment is *degenerate* if its length $m_i < 2$; degenerate segments receive a placeholder score of 0 and are excluded from C_{mag} aggregation via a validity mask. The placeholder value of 0 does not represent a valid neutral forecast.

All-degenerate windows. If no nondegenerate forecast segments remain, no valid risk score is assigned, and the window is excluded from risk-based evaluation.

B.7 Scoring Implementation Conventions

The C_{mag} and C_{risk} formulae are defined in the main paper and not repeated here. The following implementation conventions supplement those definitions.

Main setting: $\alpha = 0.75$.

Direction-mismatch guard: The directional penalty is applied only when *both* s_{text} and s_{val} are non-zero and have opposite signs. When either signal is neutral (score $= 0$), the penalty is not triggered.

Four-agent requirement: For CoRE-Ensemble, a backbone–window instance must have valid scores from all four Agent A models; windows with fewer than four valid agent scores are excluded from aggregate metric computation. The four valid risk scores are averaged before ranking, rather than averaging their evaluation metrics.

C Evaluation Guards and Valid Counts

This appendix specifies the tie treatment, metric eligibility, aggregation, and uncertainty estimation used in Section 4.2.

C.1 Tie-Aware Selective Evaluation

Evaluation is performed within each configuration–backbone group. Suppressing these indices, let n denote the number of eligible windows and e_j their window-level MAE. Windows are retained in increasing risk order. At target coverage $q \in (0, 1]$, the retained count and actual coverage are

$$k(q) = \max(1, \lceil qn \rceil), \qquad \widetilde{q}(q) = \frac{k(q)}{n}.$$

When the cutoff falls within a tied group, retained error is evaluated as the expectation under a uniform random ordering within that group. Let $\mathcal{L}_k$ contain the windows with risk strictly below the cutoff score and $\mathcal{T}_k$ those tied at the cutoff. Writing $m_k = |\mathcal{L}_k|$ and $t_k = |\mathcal{T}_k|$, the expected retained MAE is

$$E(k) = \frac{1}{k}\left[\sum_{j\in\mathcal{L}_k} e_j + \frac{k - m_k}{t_k}\sum_{j\in\mathcal{T}_k} e_j\right].$$

Each window tied at the cutoff therefore receives the same expected inclusion weight. Neither input order nor realized error determines which tied windows are preferentially retained. Tie membership uses exact equality of the stored floating-point risk scores, without a tolerance. Permuting instances within a tied group does not change the expected retained MAE.

Relative retained error and percentage improvement are

$$R(k/n) = \frac{E(k)}{\bar{e}}, \qquad I(q) = 100\left[1 - R(\widetilde{q}(q))\right], \qquad \bar{e} = \frac{1}{n}\sum_{j=1}^{n} e_j.$$

Provided $\bar{e} > 0$, $R(1) = 1$. Positive $I(q)$ indicates improvement over full coverage, whereas negative values indicate degradation. We use I_{70} and I_{80} to denote $I(0.7)$ and $I(0.8)$, respectively.

Because retained counts are integers, actual coverage can exceed target coverage. For a configuration with 24 windows, target coverages of 100%, 85%, and 70% retain 24, 21, and 17 windows, giving actual coverages of 100%, 87.5%, and approximately 70.8%. Detailed MAE tables report $E(k(q))$ together with $I(q)$. For aggregate curves, the same target coverage is applied within each group before aggregation, even when actual retained fractions differ between groups.

Figure 4(a) displays target coverages from 0.50 to 1.00 in increments of 0.05. This display grid does not change the integration grid used for U. Panel (b) uses differences in I_{70} between CoRE-Ensemble and each baseline.

C.2 Normalized Selective Utility and Reference Orderings

Let $q_k = k/n$ for $k = 1, \ldots, n$. At every retained count k, $R(q_k)$ is computed from the tie-aware expected retained MAE defined above. Thus, risk ties are averaged over possible within-tie orderings throughout the curve, not only at selected target coverages.

The area under each relative retained-error curve is

$$A = \sum_{k=1}^{n-1} \frac{R(q_k) + R(q_{k+1})}{2}(q_{k+1} - q_k).$$

Integration uses the complete attainable grid, without adding a point at zero, irrespective of the coverage range displayed in a figure. The exact expected Random curve is $R^{\mathrm{rand}}(q_k) = 1$, so $A^{\mathrm{rand}} = (n-1)/n$. Oracle orders windows by increasing realized error and uses the same integration rule.

Normalized selective utility is

$$U = \frac{A^{\text{rand}} - A^{\text{risk}}}{A^{\text{rand}} - A^{\text{oracle}}}.$$

When the relative-error normalization is valid and $A^{\text{rand}} - A^{\text{oracle}} > 0$, $U = 0$ denotes random-level utility, $U = 1$ denotes oracle-level utility, and negative values indicate performance worse than random. The normalization does not impose zero as a lower bound. Negative values are retained in aggregation without clipping, and a positive macro-average does not imply improvement in every configuration–backbone group.

For constant risk scores, tie-aware retention gives $R(q_k) = 1$ at every retained count and hence $U = 0$ whenever the normalization denominators are valid. The corresponding Spearman correlation remains undefined.

Random is evaluated analytically rather than from a sampled permutation; its reported utility and fixed-coverage improvement are both zero. Oracle uses realized targets only as an evaluation reference and is not an available scoring method at forecast time.

C.3 Nested Macro Aggregation

We assign equal weight to configurations within each backbone and family, to backbones within each family, and finally to families. Let $\mathcal{F}$ denote eligible families, $\mathcal{B}_f$ the eligible backbones in family f, and $\mathcal{C}_{f,b}$ the eligible configurations for backbone b in family f. For a configuration-level metric $M_{c,b}$, define

$$M_{f,b} = \frac{1}{|\mathcal{C}_{f,b}|} \sum_{c \in \mathcal{C}_{f,b}} M_{c,b}, \qquad M_f = \frac{1}{|\mathcal{B}_f|} \sum_{b \in \mathcal{B}_f} M_{f,b},$$

$$M_{\text{macro}} = \frac{1}{|\mathcal{F}|} \sum_{f \in \mathcal{F}} M_f.$$

Eligible sets are defined for the method and metric being evaluated. Undefined values are excluded, and any resulting differences in aggregation support are reported in the eligibility diagnostics.

This hierarchy prevents families with more configurations from dominating the macro-average. It is applied to ρ, U, $I(q)$, and relative retained error at each displayed target coverage. For per-backbone summaries, $M_{f,b}$ is averaged over eligible families without averaging over backbones. Single-agent results apply the same aggregation separately to each agent's risk ordering.

Relative errors are normalized by the full-coverage MAE within each configuration–backbone group before aggregation. When the same eligible groups and weights are used, this gives

$$I_{\text{macro}}(q) = 100\,[1 - R_{\text{macro}}(q)]\,,$$

where $R_{\text{macro}}(q)$ aggregates relative retained errors obtained by applying target coverage q in each group. Raw MAE values in the detailed table remain specific to their configuration and backbone rather than being pooled across datasets.

We do not compute a pooled Spearman correlation over windows from different configurations, because it would mix within-configuration risk ordering with differences in error scale across configurations.

C.4 Metric-Specific Guards

All methods in a direct comparison are evaluated on the same eligible backbone–window instances. The full-coverage error and the Random and Oracle references are computed on that comparison cohort. Undefined metrics are distinguished from valid zero values. Risk scores and realized errors must be finite for included instances. Non-finite risk–error pairs are removed at the window level. After filtering, groups with fewer than two instances are excluded from all metric aggregates. This minimum applies uniformly to all five TSFMs and both LLM-based forecasters, including TiRex. Under the primary BGE scoring path, an instance with no nondegenerate forecast segments has no valid risk score. CoRE-Ensemble additionally requires valid scores from all four agents; incomplete instances are not assigned a substitute risk score.

Table 8: Primary-cohort metric eligibility. Instances count finite risk–error pairs; groups are configuration–backbone pairs. Single-agent rows are evaluated separately for each named Agent A.

Method	Instances	Groups (ρ)	Groups (U)	Families
HistWidth	8,890	405	405	22
PIDEV	8,890	405	405	22
History-Trend	8,890	405	405	22
Pairwise-Point	8,890	405	405	22
CoRE-Ensemble	8,890	405	405	22
CoRE / Gemini	8,890	405	405	22
CoRE / GPT-6	8,890	400	405	22
CoRE / Fable	8,890	400	405	22
CoRE / DeepSeek	8,890	400	405	22
Pairwise-Ord	8,841	405	405	22

Spearman correlation requires at least two observations and non-constant risk and error values. Ties receive average ranks. A configuration with constant risk or constant error has undefined Spearman correlation and is excluded from the correlation aggregate; its correlation is not replaced by zero. Constant risk means that all risk values are identical; no standard-deviation threshold is imposed. Any non-finite Spearman result is treated as undefined.

Relative retained error requires a positive full-coverage mean error, $\bar{e} > 0$. The same requirement applies to percentage improvement. Normalized selective utility additionally requires valid integrated curves and a positive normalization denominator:

$$A^{\mathrm{rand}} - A^{\mathrm{oracle}} > 0.$$

If all window errors are equal, the Random and Oracle curves coincide and U is undefined, including when the common error is positive. If all errors are zero, relative retained error is also undefined. These guards use the direct comparisons $\bar{e} > 0$ and $A^{\mathrm{rand}} > A^{\mathrm{oracle}}$, without an additional positive tolerance.

Constant risk alone does not invalidate selective evaluation. Under tie-aware retention, it gives $E(k) = \bar{e}$ at every retained count, and hence $R = 1$ and $I = 0$ when $\bar{e} > 0$. It also gives $U = 0$ when the utility denominator is valid. The Random reference's reported $\rho = 0$ denotes the expected correlation of an independent random ranking, rather than the Spearman correlation of a constant risk vector.

Metric eligibility is assessed separately: a configuration with undefined correlation may still contribute to selective metrics. Any differences in valid aggregation support must be reported alongside the affected comparison.

C.5 Valid Windows, Configurations, and Families

The primary cohort contains 1,778 windows from 81 configurations across 22 dataset families. Evaluation with five forecasting backbones yields 8,890 backbone–window instances and 405 configuration–backbone groups.

The canonical eligibility audit confirms that all five methods in Table 1 retain the complete primary cohort. For each method, all 405 groups contribute valid ρ and U values, and all 22 families contribute to both macro summaries. Thus, differences among these methods are not attributable to metric-specific exclusions or differing aggregation support.

For single-agent CoRE with GPT-6 Astra, Fable 5.1 Max, or DeepSeek V4 Pro, five groups per agent have constant risk and are excluded only from the correlation aggregate. All 405 groups remain valid for U. Gemini 3.6 Flash has no such correlation exclusions. All 22 families remain represented for every agent and both metrics.

Pairwise-Ord, the additional ordinal pairwise variant, has 49 missing scores across 14 configuration–backbone groups, leaving 8,841 finite instances. Each affected group retains at least two instances,

Table 9: Matched-cohort eligibility for CoRE-Ensemble and each single-agent CoRE variant. Every row applies separately to the ensemble and to each of the four agents; counts are not summed across variants.

Backbones	Instances	Groups (ρ)	Groups (U)	Families
Five primary TSFMs	4,455	195	195	22
LLMTime	891	39	39	22
LSTPrompt	891	39	39	22

so all 405 groups remain valid for both metrics. These counts describe the available support for Pairwise-Ord, which is excluded from the main baseline comparison of 8,890 complete instances.

The matched short-horizon cohort contains 891 windows from 39 configurations across 22 families. Both CoRE-Ensemble and each of the four single-agent variants have complete eligibility, as shown in Table 9. The five TSFMs contribute 195 groups, and each LLM-based forecaster contributes 39 groups. Together, the seven backbones yield 6,237 instances and 273 groups per CoRE variant. The minimum group size is seven; there are no constant-risk exclusions. Matched results are reported separately from primary results.

C.6 Family-Bootstrap Confidence Intervals

We first compute the 22 family-level values using the aggregation defined above. Each bootstrap replicate samples 22 entries from this array with replacement and computes their mean. Repeatedly sampled families retain their multiplicity. We use $B = 10{,}000$ replicates and report the 2.5th and 97.5th percentiles.

For a paired comparison of methods A and B, we first compute their difference within each family on the common evaluation cohort:

$$\Delta_f = M_f^A - M_f^B.$$

We then resample the 22 paired differences using the same family-bootstrap procedure. A positive mean difference whose interval includes zero is not interpreted as a statistically resolved advantage. Confidence intervals for individual methods do not substitute for confidence intervals on paired differences.

For risk–coverage curves, this procedure is applied to family-level relative retained errors at each displayed target coverage. The resulting intervals are pointwise rather than simultaneous confidence bands. For per-backbone or single-agent summaries, bootstrap resampling uses the corresponding family-level metric array.

D RELIABILITY BASELINE DEFINITIONS

We compare CoRE with four alternative reliability signals: historical variability, history–forecast deviation, past-trend consistency, and numerical forecast disagreement. All methods assign risk to the same fixed target forecasts, with larger scores indicating greater predicted risk. The primary comparison uses 8,890 backbone–window instances. Evaluation, tie handling, and hierarchical aggregation follow Appendix C.

Let $\mathbf{x} = (x_1, \ldots, x_T)$ be the observed context and H the forecast horizon. Backbone b produces the target forecast $\hat{\mathbf{y}}_b = (\hat{y}_{b,1}, \ldots, \hat{y}_{b,H})$. We define the robust context scale as

$$\begin{aligned} m_x &= \mathrm{median}_{t=1,\ldots,T}\, x_t, \\ \mathrm{MAD}(\mathbf{x}) &= \mathrm{median}_{t=1,\ldots,T}\, |x_t - m_x|, \\ S_x &= \max\left\{1.4826\ \mathrm{MAD}(\mathbf{x}), 10^{-6}(|m_x|+1), 10^{-8}\right\}. \end{aligned}$$

This scale is computed from the full observed context. We set $\varepsilon = 10^{-8}$ wherever an additional denominator stabilizer is used below.

D.1 HISTWIDTH

HistWidth uses the width of a history-derived bootstrap interval as its risk signal. It tests whether historical variability alone provides a useful ordering of forecast errors.

History-derived interval construction. We retain the most recent

$$L = \min\{T, \max(128, 4H)\}$$

observations and index them as $\tilde{x}_t = x_{T-L+1+t}$ for $t = 0, \ldots, L-1$. A Theil–Sen linear fit provides an intercept $\hat{a}$ and slope $\hat{b}$. The detrended deviations and their median-centered values are

$$r_t = \tilde{x}_t - (\hat{a} + \hat{b}t), \qquad r_t^{\circ} = r_t - \mathrm{median}_{u=0,\ldots,L-1}\, r_u.$$

We generate $B_{\mathrm{hist}} = 1{,}000$ bootstrap paths using a circular moving-block bootstrap with block length

$$\ell = \max\{2, \lceil \sqrt{L} \rceil\}.$$

For each path, block starting positions are sampled uniformly from the circular sequence of centered deviations. Sampled blocks are concatenated and truncated to length H, yielding $\tilde{r}_{j,h}$ for path j and forecast step h. The bootstrap continuations are

$$\tilde{y}_{j,h} = \hat{a} + \hat{b}(L - 1 + h) + \tilde{r}_{j,h}, \qquad h = 1, \ldots, H.$$

For each forecast step, we compute the empirical quantiles

$$L_h = Q_{0.10}(\tilde{y}_{\cdot,h}), \qquad M_h = Q_{0.50}(\tilde{y}_{\cdot,h}), \qquad U_h = Q_{0.90}(\tilde{y}_{\cdot,h}).$$

These quantities depend on the observed context and horizon, and are shared across target backbones for the same window.

Risk score. HistWidth averages the interval widths after normalization by the context scale:

$$s_{\mathrm{HW}} = \frac{1}{H} \sum_{h=1}^{H} \frac{U_h - L_h}{S_x + \varepsilon}. \tag{2}$$

The score is backbone-independent. A larger value indicates a wider bootstrap interval relative to historical scale. A negative interval width is treated as invalid rather than converted into a risk score. Zero widths remain valid because the denominator is positive.

HistWidth is a study-specific baseline informed by the use of interval width to summarize distributional concentration (Gneiting et al., 2007). It applies that idea to history-derived bootstrap continuations rather than directly reproducing a published reliability estimator. The resulting intervals are not assumed to provide calibrated coverage of future targets.

D.2 PIDEV

PIDEV measures the deviation of the target point forecast from the bootstrap median, normalized by the interval half-width. It uses the same L_h, M_h, and U_h constructed for HistWidth:

$$s_{\mathrm{PIDEV},b} = \frac{1}{H}\sum_{h=1}^{H} \frac{|\hat{y}_{b,h} - M_h|}{\frac{1}{2}(U_h - L_h) + \varepsilon}. \tag{3}$$

Unlike HistWidth, this score is backbone-specific because it depends on the target forecast.

A score of zero means that the forecast coincides with the bootstrap median at every step. Positive scores can occur even when the forecast lies inside the interval. The score therefore measures standardized deviation from the median, rather than only excursions beyond the interval boundaries. Moreover, M_h need not equal $(L_h + U_h)/2$, so a score of one does not generally identify an interval boundary. For a zero-width interval, the denominator reduces to ε.

PIDEV is a study-specific baseline that tests whether departure from a history-derived reference trajectory is sufficient to identify unreliable forecasts. Interval scoring provides related background on assessing deviation relative to interval bounds (Gneiting and Raftery, 2007), but PIDEV is not a direct reproduction of a proper interval score. It compares a point forecast with a historical reference, and no proper-scoring or calibration guarantee is claimed.

D.3 HISTORY-TREND

History-Trend replaces CoRE's textual future outlook with a single ordinal signal computed from recent history. It tests whether agreement with past trend can provide the reliability information captured by the textual path.

Historical trend signal. We use the last $n = \min(H, T)$ observations:

$$x_t^{\mathrm{tail}} = x_{T-n+1+t}, \qquad t = 0, \ldots, n-1.$$

An ordinary least-squares fit against the integer time index gives slope $\hat{\beta}$. We normalize it by the absolute mean of the same tail:

$$\bar{x}_{\mathrm{tail}} = \frac{1}{n}\sum_{t=0}^{n-1} x_t^{\mathrm{tail}}, \qquad z_{\mathrm{hist}} = \frac{\hat{\beta}}{|\bar{x}_{\mathrm{tail}}| + \varepsilon}.$$

This historical window differs from the longer tail used to construct HistWidth and PIDEV.

The relative slope is mapped to a seven-level signal $s_{\mathrm{hist}} = Q(z_{\mathrm{hist}})$, where

$$Q(z) = \begin{cases} 3, & z \geq 0.20, \\ 2, & 0.08 \leq z < 0.20, \\ 1, & 0.02 \leq z < 0.08, \\ 0, & -0.02 \leq z < 0.02, \\ -1, & -0.08 < z < -0.02, \\ -2, & -0.20 < z \leq -0.08, \\ -3, & z \leq -0.20. \end{cases}$$

In particular, $z = -0.08$ maps to -2 and $z = -0.20$ maps to -3. A single historical level is used as the reference for all future segments; the historical tail is not divided into three independently scored reference segments.

Consistency and risk. The target forecast uses the same $K = 3$ future segments and numerical ordinal signals $s_{\mathrm{val}}^{(k)}$ as CoRE. Let $\mathcal{V}$ be the set of nondegenerate forecast segments. For each $k \in \mathcal{V}$, define

$$b_k^{\mathrm{hist}} = 1 - \frac{|s_{\mathrm{hist}} - s_{\mathrm{val}}^{(k)}|}{6},$$

$$p_k = \begin{cases} \alpha b_k^{\mathrm{hist}}, & s_{\mathrm{hist}} s_{\mathrm{val}}^{(k)} < 0, \\ b_k^{\mathrm{hist}}, & \text{otherwise}, \end{cases} \qquad \alpha = 0.75.$$

The additional penalty applies only to opposite nonzero directions. For $\mathcal{V} \neq \emptyset$, historical consistency is

$$C_{\text{mag}}^{\text{hist}} = \frac{1}{|\mathcal{V}|} \sum_{k \in \mathcal{V}} p_k. \tag{4}$$

The risk score is

$$s_{\text{HT},b} = \begin{cases} 1 - C_{\text{mag}}^{\text{hist}}, & \mathcal{V} \neq \emptyset, \\ 0.5, & \mathcal{V} = \emptyset. \end{cases} \tag{5}$$

The value 0.5 is also used when the observed context is unavailable. This is a baseline-specific fallback and is distinct from the handling of unavailable CoRE scores.

History-Trend is a study-specific baseline derived from the CoRE consistency framework, rather than a direct reproduction of a published trend-forecasting method. Its comparison with CoRE isolates the replacement of the textual outlook by this particular historical trend representation. It does not test every possible numerical representation of observed history.

D.4 PAIRWISE-POINT

Pairwise-Point measures disagreement between the target point forecast and auxiliary numerical forecasts for the same window and horizon. Let $\mathcal{A}_b$ denote the auxiliary set for target backbone b. The score is

$$s_{\text{PW},b} = \frac{1}{|\mathcal{A}_b|} \sum_{a \in \mathcal{A}_b} \frac{1}{H S_x} \sum_{h=1}^{H} |\hat{y}_{b,h} - \hat{y}_{a,h}|. \tag{6}$$

Each auxiliary contributes equally. The scale S_x is computed from the observed context as defined above and is strictly positive.

The primary auxiliary pool consists of Chronos, TimeMoE, TimesFM, Moirai, and TiRex. For a primary TSFM target, the target backbone is excluded, leaving four auxiliaries and $|\mathcal{A}_b| = 4$. For LLMTime and LSTPrompt, all five primary TSFM backbones serve as auxiliaries, giving $|\mathcal{A}_b| = 5$. The distance calculation and scale normalization are otherwise unchanged. No auxiliary is selected according to realized forecast error.

The target forecast used in this distance calculation is the same stored forecast used to compute CoRE risk and realized MAE for that backbone–window instance. Auxiliary forecasts are also held fixed. Pairwise-Point assigns a reliability score without replacing the target prediction by an ensemble mean.

This baseline adapts the ensemble-disagreement principle used in predictive uncertainty estimation (Lakshminarayanan et al., 2017). It differs from a direct implementation of Deep Ensembles: it uses target-to-auxiliary absolute distances between heterogeneous forecasting backbones, normalized by historical scale, rather than predictive variance from a trained probabilistic ensemble. The comparison tests numerical forecast disagreement as an alternative to CoRE's cross-modal agreement.

E Detailed Reliability and Selective Forecasting Results

E.1 Per-Backbone CoRE-Ensemble Performance

Table 10: Per-backbone CoRE-Ensemble performance on the primary and matched short-horizon ($H \leq 60$) cohorts. I_{80} denotes MAE improvement over full coverage at 80% target coverage. Brackets show available 95% family-bootstrap confidence intervals. The matched cohort contains 891 windows per backbone.

	ρ_{macro}		U_{macro}		I_{80} (%)	
Backbone	Mean	95% CI	Mean	95% CI	Mean	95% CI
(a) Primary cohort: 1,778 windows, 81 configurations, 22 families						
Chronos	0.383	[0.251, 0.511]	0.410	[0.251, 0.565]	18.8	[10.6, 28.3]
TimeMoE	0.430	[0.293, 0.561]	0.464	[0.296, 0.621]	21.4	[12.6, 31.3]
TimesFM	0.379	[0.249, 0.505]	0.404	[0.260, 0.551]	15.5	[8.5, 24.0]
Moirai	0.418	[0.275, 0.552]	0.426	[0.222, 0.602]	16.3	[9.0, 24.6]
TiRex	0.370	[0.231, 0.498]	0.396	[0.230, 0.555]	15.5	[7.5, 25.1]
Cross-backbone macro	0.396	[0.266, 0.521]	0.420	[0.269, 0.567]	17.5	—
(b) Matched short-horizon cohort: 39 configurations, 22 families						
Chronos	0.350	[0.217, 0.472]	0.377	[0.206, 0.540]	20.7	[11.3, 31.5]
TimeMoE	0.408	[0.269, 0.531]	0.449	[0.279, 0.606]	22.2	[13.4, 31.8]
TimesFM	0.350	[0.215, 0.473]	0.362	[0.179, 0.529]	16.4	[8.7, 25.5]
Moirai	0.377	[0.232, 0.509]	0.408	[0.199, 0.584]	16.2	[8.9, 24.5]
TiRex	0.336	[0.195, 0.457]	0.367	[0.189, 0.530]	15.4	[6.6, 26.2]
LLMTime	0.358	—	0.367	[0.194, 0.529]	13.7	[8.3, 20.4]
LSTPrompt	0.356	—	0.342	[0.161, 0.510]	19.2	[11.8, 27.3]

Table 10 reports CoRE-Ensemble performance separately for each forecasting backbone. The ensemble averages the four Agent A risk scores before ranking. Although textual outlooks can be reused across backbones, each backbone's numerical forecast determines its own consistency score and realised error. Within each backbone, configuration-level metrics are averaged within families and then equally across families, following Appendix C.

Primary cohort. The five TSFM backbones are evaluated on 1,778 windows across 81 configurations and 22 families. Their macro-correlations range from 0.370 to 0.430, and their selective utilities range from 0.396 to 0.464. TimeMoE has the highest point estimates for both metrics. The overall primary-cohort results are $\rho_{\text{macro}} = 0.396$, $U_{\text{macro}} = 0.420$, and $I_{80} = 17.47\%$. Positive backbone-level summaries do not imply positive association or selective improvement in every configuration.

Matched short-horizon cohort. The matched evaluation covers 39 configurations and 22 families with $H \leq 60$. The seven backbones each contribute 891 windows.

Across the seven backbones, selective utility ranges from 0.342 to 0.449. LLMTime and LSTPrompt achieve $U_{\text{macro}} = 0.367$ and 0.342, respectively. Their corresponding I_{80} values are 13.7% and 19.2%. These results describe reliability ranking within each evaluated backbone; they do not establish differences in forecasting accuracy between backbones. Primary and matched results are reported separately because their window and horizon support differ.

E.2 Single-Agent CoRE Performance

Tables 11 and 12 report CoRE scores constructed separately from each Agent A model. Each single-agent score compares that agent's BGE-mapped textual outlook with the numerical forecast. Metrics follow the aggregation and uncertainty procedure in Appendix C.

Primary cohort. All 20 combinations of four agents and five TSFM backbones have positive macro-correlation and utility point estimates. Macro-correlation ranges from 0.289 to 0.410, and

Table 11: Single-agent per-backbone reliability metrics (primary cohort). Each cell corresponds to one Agent A model evaluated against one TSFM backbone's point forecasts using the canonical tie-aware risk–coverage protocol (Appendix C). Primary cohort: 1,778 windows, 81 configurations, 22 families, five zero-shot TSFM backbones; aggregation: config → family (equal weight, 22 families). ρ_{macro}: Spearman rank correlation (risk vs. MAE). U_{macro}: normalized selective utility (tie-aware). 95% bootstrap CIs ($B = 10{,}000$) in brackets. The cross-backbone macro row averages the five backbone-level family means for each agent.

Agent A	Backbone	ρ_{macro}	ρ 95% CI	U_{macro}	U 95% CI
Gemini 3.6 Flash	Chronos	0.318	[0.188, 0.447]	0.341	[0.166, 0.506]
	TimeMoE	0.335	[0.181, 0.484]	0.386	[0.208, 0.556]
	TimesFM	0.317	[0.182, 0.453]	0.287	[0.109, 0.466]
	Moirai	0.357	[0.215, 0.495]	0.374	[0.170, 0.554]
	TiRex	0.312	[0.169, 0.448]	0.326	[0.138, 0.500]
	Cross-backbone macro	*0.327*	—	*0.343*	—
GPT-6 Astra	Chronos	0.294	[0.165, 0.419]	0.323	[0.160, 0.476]
	TimeMoE	0.350	[0.230, 0.468]	0.375	[0.227, 0.519]
	TimesFM	0.294	[0.177, 0.413]	0.314	[0.176, 0.456]
	Moirai	0.335	[0.206, 0.458]	0.323	[0.143, 0.484]
	TiRex	0.289	[0.160, 0.408]	0.319	[0.161, 0.461]
	Cross-bb macro	*0.312*	—	*0.331*	—
Fable 5.1 Max	Chronos	0.364	[0.239, 0.487]	0.367	[0.198, 0.532]
	TimeMoE	0.410	[0.281, 0.530]	0.450	[0.299, 0.592]
	TimesFM	0.338	[0.204, 0.469]	0.372	[0.219, 0.525]
	Moirai	0.374	[0.235, 0.507]	0.391	[0.211, 0.548]
	TiRex	0.356	[0.228, 0.480]	0.370	[0.193, 0.539]
	Cross-bb macro	*0.369*	—	*0.390*	—
DeepSeek V4 Pro	Chronos	0.349	[0.237, 0.456]	0.406	[0.288, 0.522]
	TimeMoE	0.384	[0.264, 0.500]	0.424	[0.275, 0.568]
	TimesFM	0.310	[0.187, 0.428]	0.354	[0.227, 0.479]
	Moirai	0.365	[0.240, 0.485]	0.382	[0.224, 0.524]
	TiRex	0.317	[0.196, 0.433]	0.373	[0.244, 0.497]
	Cross-bb macro	*0.345*	—	*0.388*	—

utility ranges from 0.287 to 0.450. The reported confidence intervals for both metrics exclude zero. These ranges describe separate agent–backbone combinations; they are not uncertainty intervals for CoRE-Ensemble.

Matched short-horizon cohort. The matched results include 28 combinations of four agents and seven forecasting backbones. LLMTime and LSTPrompt use the same forecast sources and available windows as their ensemble results in Table 10. The eight combinations involving these two LLM-based forecasters have positive utility point estimates, ranging from 0.228 to 0.383. Comparisons with primary-cohort results should account for the different evaluated windows and horizons.

Metric eligibility. Undefined correlations are excluded rather than replaced with zero. In the primary cohort, Gemini contributes all 405 configuration–backbone groups to both metrics. Each of the other three agents has five constant-risk groups, which are excluded from correlation but remain eligible for utility. All four agents therefore contribute 405 groups to utility. Metric-specific eligibility follows Appendix C.

E.3 Single-Agent versus Ensemble Comparison

Table 13 compares the four single-agent methods with CoRE-Ensemble on the same 8,890 primary-cohort backbone–window instances. The ensemble averages the four risk scores for each instance before ranking; it does not average their rankings or modify the underlying forecasts.

Aggregate performance. CoRE-Ensemble achieves $\rho_{\text{macro}} = 0.396$ with a 95% confidence interval of $[0.266, 0.521]$, and $U_{\text{macro}} = 0.420$ with an interval of $[0.269, 0.567]$. The single-agent

Table 12: Single-agent per-backbone reliability metrics (matched short-horizon cohort). Matched cohort ($H \leq 60$): 39 configurations, 22 families, five zero-shot TSFM backbones plus two LLM-adapted backbones; aggregation: config → family (equal weight, 22 families). 95% bootstrap CIs ($B = 10{,}000$) in brackets; ρ CIs not reported for LLM backbones (two-level aggregation).

Agent A	Backbone	ρ_{macro}	ρ 95% CI	U_{macro}	U 95% CI
Gemini 3.6 Flash	Chronos	0.293	[0.158, 0.420]	0.319	[0.134, 0.489]
	TimeMoE	0.316	[0.172, 0.452]	0.373	[0.197, 0.538]
	TimesFM	0.281	[0.145, 0.414]	0.229	[0.034, 0.420]
	Moirai	0.321	[0.178, 0.457]	0.357	[0.146, 0.544]
	TiRex	0.268	[0.127, 0.399]	0.285	[0.090, 0.462]
	LLMTime	0.322	—	0.364	[0.189, 0.523]
	LSTPrompt	0.300	—	0.228	[0.010, 0.435]
GPT-6 Astra	Chronos	0.247	[0.100, 0.386]	0.299	[0.109, 0.477]
	TimeMoE	0.313	[0.177, 0.440]	0.362	[0.197, 0.522]
	TimesFM	0.266	[0.130, 0.396]	0.297	[0.121, 0.467]
	Moirai	0.285	[0.151, 0.415]	0.303	[0.115, 0.472]
	TiRex	0.242	[0.097, 0.375]	0.302	[0.129, 0.465]
	LLMTime	0.270	—	0.291	[0.128, 0.453]
	LSTPrompt	0.285	—	0.272	[0.075, 0.455]
Fable 5.1 Max	Chronos	0.300	[0.168, 0.424]	0.323	[0.147, 0.491]
	TimeMoE	0.344	[0.212, 0.463]	0.413	[0.254, 0.566]
	TimesFM	0.301	[0.166, 0.423]	0.338	[0.164, 0.497]
	Moirai	0.308	[0.164, 0.440]	0.346	[0.162, 0.503]
	TiRex	0.289	[0.155, 0.409]	0.332	[0.142, 0.506]
	LLMTime	0.331	—	0.345	[0.177, 0.505]
	LSTPrompt	0.338	—	0.364	[0.181, 0.521]
DeepSeek V4 Pro	Chronos	0.340	[0.238, 0.436]	0.414	[0.290, 0.534]
	TimeMoE	0.379	[0.267, 0.483]	0.451	[0.302, 0.592]
	TimesFM	0.305	[0.187, 0.415]	0.349	[0.185, 0.501]
	Moirai	0.363	[0.241, 0.479]	0.408	[0.240, 0.554]
	TiRex	0.303	[0.193, 0.403]	0.385	[0.250, 0.511]
	LLMTime	0.331	—	0.383	[0.232, 0.522]
	LSTPrompt	0.338	—	0.358	[0.193, 0.507]

point estimates range from 0.312 to 0.369 for correlation and from 0.331 to 0.390 for utility. The ensemble has the highest observed mean for both metrics.

The retained-error ratios are normalized within each configuration–backbone group before aggregation. At 80% target coverage, the ensemble has $R_{80} = 0.825$, compared with 0.850–0.871 for the individual agents. At 70% target coverage, its ratio is 0.742, compared with 0.767–0.801. Lower ratios indicate lower retained MAE relative to full coverage.

Paired utility differences. The ensemble's utility differences from Gemini and GPT-6 Astra are $+0.077$ and $+0.089$, with respective 95% paired confidence intervals of $[0.021, 0.138]$ and $[0.011, 0.163]$. The differences from Fable and DeepSeek are $+0.030$ and $+0.032$, with intervals of $[-0.013, 0.068]$ and $[-0.034, 0.094]$. Thus, the higher ensemble mean is statistically resolved for two comparisons but not for all four. These are individual comparison intervals without multiple-comparison adjustment. The results do not establish that four agents are optimal or that ensembling improves every agent–backbone configuration.

E.4 Family-Level Reliability Results

Tables 14 and 15 report family–backbone results for the primary and matched cohorts. Each cell contains selective utility $U_{f,b}$ and full-coverage-relative improvement $I_{80,f,b}$, averaged equally over eligible configurations within that family and backbone. The bottom row averages equally over the 22 families.

Table 13: CoRE-Ensemble and single-agent results on the primary cohort. (a) Macro-averaged correlation, selective utility, and relative retained MAE at 85%, 80%, and 70% target coverage. (b) Paired utility differences, ensemble minus individual agent. Brackets report 95% family-bootstrap confidence intervals for ρ, U, and ΔU. † indicates that the comparison-wise interval excludes zero; no multiple-comparison adjustment is applied.

(a) Reliability metrics

Method	ρ_{macro}	ρ 95% CI	U_{macro}	U 95% CI	R_{85}	R_{80}	R_{70}
CoRE-Ensemble	0.396	[0.266, 0.521]	0.420	[0.269, 0.567]	0.862	0.825	0.742
Gemini 3.6 Flash	0.327	[0.193, 0.462]	0.343	[0.175, 0.505]	0.878	0.856	0.777
GPT-6 Astra	0.312	[0.196, 0.427]	0.331	[0.193, 0.467]	0.894	0.871	0.801
Fable 5.1 Max	0.369	[0.245, 0.490]	0.390	[0.240, 0.540]	0.880	0.850	0.776
DeepSeek V4 Pro	0.345	[0.232, 0.454]	0.388	[0.267, 0.506]	0.887	0.851	0.767

(b) Paired ΔU = Ensemble − Agent

Agent	ΔU (point)	ΔU 95% CI
Gemini 3.6 Flash	$+0.0773^{\dagger}$	[+0.021, +0.138]
GPT-6 Astra	$+0.0893^{\dagger}$	[+0.011, +0.163]
Fable 5.1 Max	+0.0301	[−0.013, +0.068]
DeepSeek V4 Pro	+0.0321	[−0.034, +0.094]

Table 14: Family-level CoRE-Ensemble results on the primary cohort. Each cell reports $U_{f,b}$ / $I_{80,f,b}$ (%), averaged equally over configurations within each family–backbone group. I_{80} measures MAE improvement over full coverage at 80% target coverage.

Family	**Chronos**	**TimeMoE**	**TimesFM**	**Moirai**	**TiRex**
Each cell: $U_{f,b}$ / $I_{80,f,b}$ (%)					
Bitbrains-FS	0.944 / 86.3	0.922 / 61.1	0.941 / 77.4	0.929 / 74.3	0.966 / 86.6
Bitbrains-Rnd	0.797 / 46.2	0.842 / 55.4	0.780 / 45.9	0.684 / 31.1	0.786 / 46.3
Bizitobs-App	0.932 / 4.8	0.919 / 1.1	0.963 / 14.7	0.940 / 5.5	0.853 / −2.5
Bizitobs-L2C	0.272 / 3.9	0.376 / 6.3	0.228 / 2.3	0.470 / 4.3	0.364 / 5.7
Bizitobs-Svc	0.782 / 37.0	0.795 / 39.4	0.773 / 31.0	0.811 / 38.4	0.734 / 30.2
Electricity	0.495 / 13.6	0.590 / 18.8	0.541 / 14.9	0.638 / 16.3	0.557 / 17.5
ETT-1	0.712 / 25.0	0.715 / 18.2	0.730 / 18.0	0.783 / 21.4	0.697 / 20.5
ETT-2	0.524 / 13.1	0.596 / 14.3	0.433 / 12.0	0.536 / 10.3	0.560 / 12.5
Jena Weather	0.837 / 42.3	0.863 / 44.8	0.862 / 39.9	0.870 / 49.5	0.811 / 31.5
KDD Cup 2018	−0.107 / 13.6	0.012 / 4.1	−0.059 / −0.2	−0.041 / 10.4	−0.123 / 0.3
Loop Seattle	0.418 / 3.1	0.411 / 4.9	0.519 / 6.0	0.471 / 7.0	0.386 / 3.1
M4-Daily	0.236 / 16.3	−0.140 / 17.0	0.325 / 18.9	0.326 / 18.9	0.578 / 21.9
M4-Hourly	0.820 / 23.0	0.761 / 12.6	0.786 / 9.2	0.857 / 30.5	0.776 / 15.6
M4-Monthly	−0.142 / −7.9	0.489 / 39.0	0.058 / 2.5	−0.001 / −10.8	−0.216 / −11.0
M4-Qrtly	−0.266 / 14.5	0.791 / 52.1	0.072 / 8.6	−1.032 / 4.7	−0.222 / 4.7
M4-Weekly	0.606 / 31.1	0.850 / 67.9	0.512 / 15.6	0.602 / 8.5	0.525 / 13.7
M-Dense	0.366 / 11.9	0.319 / 7.1	0.243 / 9.4	0.544 / 12.9	0.284 / 9.4
Saugeen	0.017 / 5.4	−0.322 / 2.3	0.140 / 5.3	0.318 / 6.8	0.078 / 1.8
Solar	0.033 / −3.2	0.097 / 2.8	−0.236 / −4.7	0.059 / −2.5	−0.125 / −2.7
SZ-Taxi	0.012 / −0.6	0.030 / 0.7	0.019 / −0.6	0.064 / 2.7	−0.036 / 0.6
Temp. Rain	0.730 / 35.7	0.536 / 3.6	0.214 / 13.9	0.678 / 17.8	0.654 / 35.3
US Births	−0.0002 / −2.4	−0.249 / −3.4	0.049 / 0.5	−0.127 / −0.5	−0.167 / −0.7
Macro (22 fam.)	0.410 / 18.8	0.464 / 21.4	0.404 / 15.5	0.426 / 16.3	0.396 / 15.5

Variation across families. In the primary cohort, 19 of the 110 family–backbone cells have negative utility, and 15 have negative I_{80}. Fourteen families have positive utility across all five TSFM backbones. The positive overall macro-average therefore coexists with families in which retention provides little benefit or increases error. The matched results likewise contain both positive and negative values, which are retained without clipping.

Table 15: Family-level CoRE-Ensemble results on the matched short-horizon cohort ($H \leq 60$). Each cell reports $U_{f,b}$ / $I_{80,f,b}$ (%). Results cover 39 configurations and 22 families, with 891 windows per backbone.

Family	**Chronos**	**TimeMoE**	**TimesFM**	**Moirai**	**TiRex**	**LLMTime**	**LSTPrompt**
Each cell: U_{macro} / I_{80} (%)							
Bitbrains-FS	0.921 / 99.0	0.879 / 43.5	0.917 / 77.4	0.892 / 72.2	0.967 / 98.0	0.789 / 9.7	0.749 / 45.7
Bitbrains-Rnd	0.856 / 52.9	0.945 / 71.2	0.818 / 51.8	0.650 / 20.9	0.834 / 53.4	0.878 / 65.7	0.743 / 57.2
Bizitobs-App	0.797 / 14.5	0.758 / 3.4	0.888 / 44.1	0.819 / 16.5	0.651 / −7.4	0.794 / 11.3	0.913 / 49.1
Bizitobs-L2C	0.277 / 11.0	0.391 / 16.5	0.164 / 5.8	0.491 / 6.1	0.491 / 19.6	0.484 / 3.8	0.282 / 8.6
Bizitobs-Svc	0.816 / 39.2	0.839 / 48.6	0.787 / 22.7	0.879 / 47.8	0.754 / 28.9	0.758 / 12.5	0.809 / 41.7
Electricity	0.211 / 4.3	0.381 / 8.7	0.270 / 5.2	0.435 / 0.9	0.333 / 10.6	0.235 / −1.8	0.417 / 28.0
ETT-1	0.596 / 32.3	0.656 / 21.6	0.635 / 17.2	0.675 / 21.9	0.582 / 23.7	0.563 / 26.1	0.431 / 18.5
ETT-2	0.204 / 4.4	0.355 / 5.8	0.060 / 7.0	0.284 / 1.5	0.285 / 2.0	0.301 / 8.0	0.057 / 2.9
Jena Weather	0.755 / 24.5	0.828 / 35.3	0.856 / 40.8	0.801 / 36.5	0.706 / 10.4	0.765 / 21.7	0.570 / 8.4
KDD Cup 2018	0.565 / 40.3	0.596 / 10.3	0.512 / −1.4	0.623 / 28.2	0.468 / −1.1	0.168 / 2.5	0.355 / 1.7
Loop Seattle	0.452 / 5.7	0.451 / 8.3	0.591 / 6.6	0.493 / 5.6	0.484 / 4.2	0.423 / 4.5	0.444 / 2.7
M4-Daily	0.236 / 16.3	−0.140 / 17.0	0.325 / 18.9	0.326 / 18.9	0.578 / 21.9	0.276 / 21.7	−0.178 / 19.2
M4-Hourly	0.820 / 23.0	0.761 / 12.6	0.786 / 9.2	0.857 / 30.5	0.776 / 15.6	0.814 / 12.6	0.901 / 49.2
M4-Monthly	−0.142 / −7.9	0.489 / 39.0	0.058 / 2.5	−0.001 / −10.8	−0.216 / −11.0	0.063 / 7.4	−0.687 / −1.8
M4-Qrtly	−0.266 / 14.5	0.791 / 52.1	0.072 / 8.6	−1.032 / 4.7	−0.222 / 4.7	−0.240 / 17.2	−0.093 / 7.2
M4-Weekly	0.606 / 31.1	0.850 / 67.9	0.512 / 15.6	0.602 / 8.5	0.525 / 13.7	0.606 / 28.8	0.595 / 26.4
M-Dense	0.360 / 24.8	0.337 / 18.1	0.184 / 22.1	0.649 / 22.4	0.266 / 21.9	0.539 / 13.4	0.066 / 7.5
Saugeen	0.017 / 5.4	−0.322 / 2.3	0.140 / 5.3	0.318 / 6.8	0.078 / 1.8	−0.398 / 3.4	0.547 / 17.1
Solar	−0.536 / −13.6	−0.251 / 3.1	−0.891 / −12.3	−0.424 / −6.7	−0.657 / −8.0	−0.481 / 11.0	0.684 / 29.5
SZ-Taxi	0.022 / 0.2	0.005 / 1.8	0.029 / −0.3	0.095 / 5.7	−0.092 / 1.1	0.021 / 1.7	0.141 / 0.1
Temp. Rain	0.730 / 35.7	0.536 / 3.6	0.214 / 13.9	0.678 / 17.8	0.654 / 35.3	0.766 / 24.5	0.203 / 10.1
US Births	−0.0002 / −2.4	−0.249 / −3.4	0.049 / 0.5	−0.127 / −0.5	−0.167 / −0.7	−0.049 / −5.2	−0.431 / −6.8
Macro (22 fam.)	0.377 / 20.7	0.449 / 22.2	0.362 / 16.4	0.408 / 16.2	0.367 / 15.4	0.367 / 13.7	0.342 / 19.2

Utility and fixed-coverage improvement. Utility summarizes the complete attainable coverage range, whereas I_{80} describes one operating point. They need not have the same sign. For example, Moirai on M4-Quarterly has $U = -1.032$ but $I_{80} = 4.7\%$ in the primary cohort. A negative utility value therefore does not imply degradation at every coverage. Conversely, positive utility does not guarantee improvement at a particular retention level.

E.5 Risk–Coverage Curves

Figures 5 and 6 present complementary views of coverage-dependent performance. Relative retained MAE is normalized within each configuration–backbone group and then aggregated using the hierarchy in Appendix C. Lower values indicate better selective performance, and the exact Random reference is $R = 1$.

Comparison within each backbone. Figure 5 compares the four single-agent methods and CoRE-Ensemble separately for each forecasting backbone. The panels distinguish the five primary TSFM backbones from the two matched LLM-based forecasters. This view shows whether differences between textual evaluators persist across coverage levels within a given backbone.

Comparison within each agent. Figure 6 groups the curves by Agent A, allowing comparison across forecasting backbones for the same textual evaluator. The TSFM and LLM-based results are displayed separately because they use different cohort and horizon support.

The displayed coverage range does not change the definition of utility, which integrates over the complete attainable grid. Curve differences describe observed retained-error behavior; they do not by themselves establish a statistically resolved advantage at every coverage.

E.6 Family-Level Selective-MAE Summaries

Tables 16 and 17 report absolute retained MAE for CoRE-Ensemble and Pairwise-Point at 100%, 85%, and 70% target coverage. The tables cover the 22 dataset families and seven forecasting backbones. TSFM results use the primary cohort, while LLMTime and LSTPrompt use their matched short-horizon evaluations.

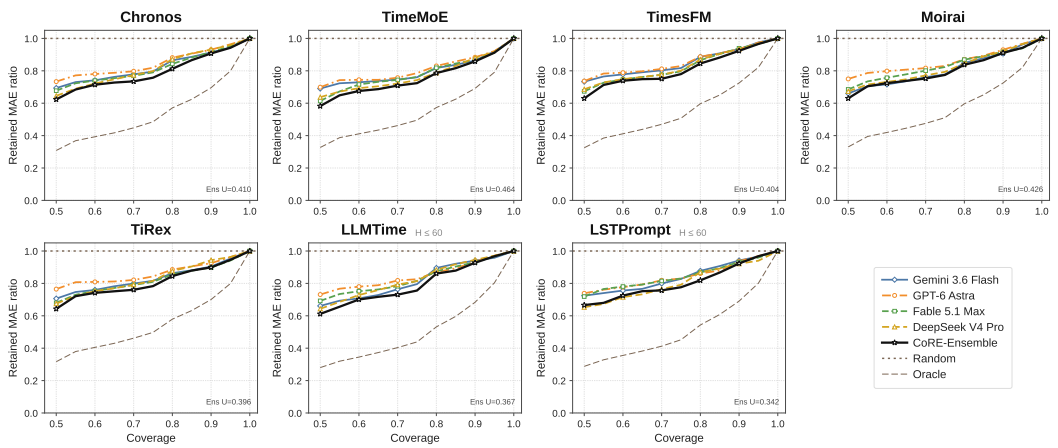


Figure 5: Risk–coverage curves for four single-agent CoRE variants and CoRE-Ensemble, grouped by forecasting backbone. The five TSFMs use the primary cohort; LLMTime and LSTPrompt use the matched short-horizon cohort. Target coverage ranges from 50% to 100%. The vertical axis reports family-macro relative retained MAE. The annotated U in each panel is the CoRE-Ensemble utility for that backbone (Table 10), computed over the complete attainable coverage grid. Random and Oracle provide reference orderings.

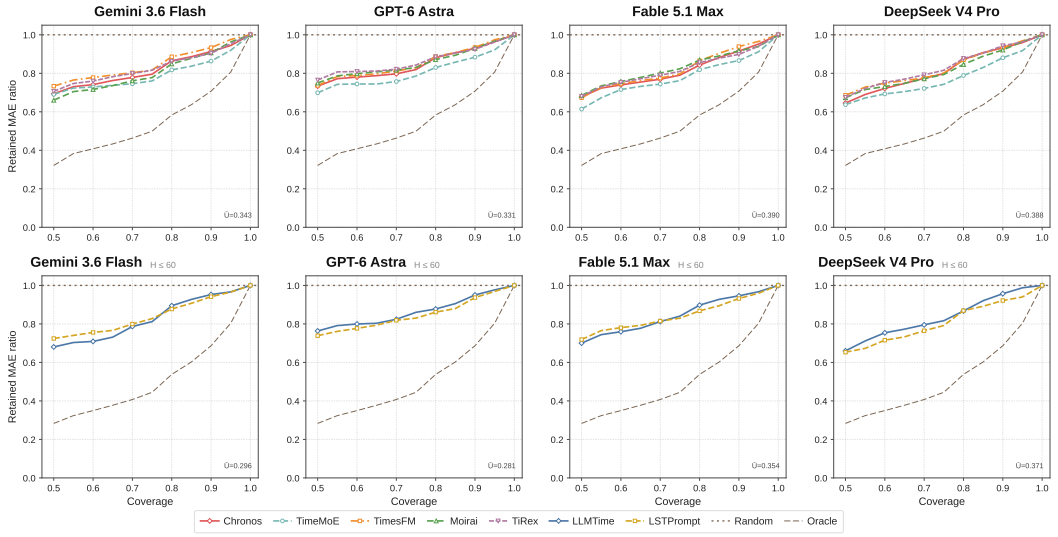


Figure 6: Risk–coverage curves for the four single-agent CoRE variants, grouped by Agent A. The upper row compares five TSFM backbones on the primary cohort; the lower row compares LLMTime and LSTPrompt on the matched short-horizon cohort. Target coverage ranges from 50% to 100%. The vertical axis reports family-macro relative retained MAE. Each $\bar{U}$ annotation is the mean single-agent utility of that Agent A over the backbones displayed in that panel. Random and Oracle provide reference orderings.

Within each backbone, both methods use the same forecast run and common eligible windows and configurations. The methods therefore have identical full-coverage MAE within each reported comparison.

For family f, backbone b, and target coverage q, define the family-level retained MAE as

$$\overline{E}_{f,b}(q) = \frac{1}{|\mathcal{C}_{f,b}|} \sum_{c \in \mathcal{C}_{f,b}} E_{c,b}(q),$$

where $\mathcal{C}_{f,b}$ is the common set of eligible configurations. The percentage improvement displayed in parentheses is

$$J_{f,b}(q) = 100 \left[1 - \frac{\overline{E}_{f,b}(q)}{\overline{E}_{f,b}(1)} \right].$$

Thus, each displayed percentage is calculated directly from the family-level MAE values in the same table. It differs from $I_{f,b}(q)$ in the preceding sections, which averages configuration-level percentage improvements.

Raw MAE retains the scale of the underlying observations. Where a family contains multiple configurations, its mean can combine different frequencies and horizons. The tables therefore support comparisons between retention methods within the same family and backbone, rather than comparisons of absolute MAE across dataset families.

Table 16: CoRE-Ensemble retained MAE at 85% and 70% target coverage by dataset family and backbone. Parentheses show percentage improvement over full-coverage family-mean MAE. TSFM results use the primary cohort; LLM-based results use the matched short-horizon cohort.

(a) MAE (Improvement, %)

Family	Chronos		TimeMoE		TimesFM		Moirai	
	85%	70%	85%	70%	85%	70%	85%	70%
Bitbrains-FS	149.65(55.4%)	11.10(96.7%)	140.19(54.6%)	9.75(96.8%)	98.67(70.5%)	11.02(96.7%)	99.66(67.7%)	51.07(83.5%)
Bitbrains-Rnd	21.94(33.6%)	12.81(61.2%)	19.13(54.0%)	16.53(60.2%)	27.55(32.8%)	18.39(55.1%)	32.13(26.4%)	26.78(38.7%)
Bizitobs-App	1898.43(3.5%)	1465.05(25.5%)	2726.46(0.8%)	2035.08(26.0%)	2708.91(10.5%)	1731.54(42.8%)	2402.74(3.9%)	1861.32(25.6%)
Bizitobs-L2C	10.59(0.1%)	9.92(6.4%)	9.52(4.2%)	9.00(9.4%)	10.41(2.2%)	10.44(1.9%)	11.32(3.7%)	9.97(15.2%)
Bizitobs-Svc	60.82(21.5%)	32.42(58.2%)	92.64(22.3%)	46.48(61.0%)	95.52(18.7%)	49.54(57.8%)	76.87(22.3%)	39.75(59.8%)
Electricity	423.15(8.6%)	327.04(29.4%)	381.31(11.9%)	279.38(35.4%)	394.19(9.4%)	296.93(31.8%)	658.00(9.2%)	456.12(37.0%)
ETT1	22.31(17.5%)	17.59(34.9%)	20.81(13.2%)	16.92(29.4%)	20.63(12.6%)	15.92(32.6%)	22.62(17.3%)	16.41(40.0%)
ETT2	29.08(9.7%)	25.60(20.5%)	34.42(10.1%)	29.70(22.4%)	35.87(9.9%)	33.15(16.7%)	38.07(8.9%)	34.94(16.4%)
Jena Weather	6.66(36.8%)	5.02(52.4%)	7.36(35.4%)	5.28(53.7%)	6.78(33.0%)	4.35(57.0%)	10.16(40.7%)	6.73(60.7%)
KDD Cup 2018	11.20(1.4%)	9.98(12.2%)	9.35(1.1%)	9.09(3.8%)	10.88(−0.6%)	10.58(2.2%)	10.80(5.4%)	10.29(9.9%)
Loop Seattle	5.74(3.2%)	5.69(4.1%)	5.50(4.2%)	5.36(6.5%)	6.54(5.3%)	6.16(10.9%)	8.13(5.4%)	7.81(9.1%)
M4-Daily	154.68(10.1%)	120.51(29.9%)	188.25(11.2%)	177.55(16.2%)	140.95(18.6%)	138.44(20.0%)	144.57(12.4%)	120.91(26.8%)
M4-Hourly	297.25(17.7%)	161.80(55.2%)	272.15(10.0%)	167.30(44.7%)	329.27(4.6%)	164.04(52.5%)	754.82(21.7%)	346.16(64.1%)
M4-Monthly	734.18(−8.3%)	791.82(−16.8%)	702.08(38.9%)	730.09(36.4%)	460.60(1.4%)	438.38(6.2%)	807.24(−7.3%)	932.36(−23.9%)
M4-Quarterly	241.71(16.0%)	256.00(11.0%)	169.81(52.5%)	145.19(59.4%)	222.25(11.1%)	218.63(12.5%)	244.21(8.4%)	278.59(−4.4%)
M4-Weekly	139.43(32.5%)	112.43(45.6%)	213.70(67.5%)	190.69(71.0%)	218.54(4.5%)	168.26(26.5%)	246.63(9.2%)	179.16(34.0%)
M-Dense	61.19(12.0%)	60.80(12.6%)	75.89(7.6%)	74.24(9.6%)	61.48(9.5%)	60.72(10.6%)	210.47(11.7%)	191.81(19.5%)
Saugeen	16.13(0.2%)	14.53(10.1%)	15.11(3.3%)	15.31(2.1%)	14.45(0.4%)	12.77(12.0%)	15.71(5.1%)	14.02(15.3%)
Solar	8.14(−3.9%)	8.11(−3.6%)	7.47(−0.8%)	7.35(0.9%)	10.33(−4.6%)	10.75(−8.8%)	19.64(−2.2%)	19.39(−0.9%)
SZ-Taxi	2.63(1.1%)	2.68(−0.6%)	2.57(1.3%)	2.65(−1.7%)	2.50(1.1%)	2.59(−2.1%)	2.83(4.5%)	2.95(0.2%)
Temp-Rain	2.22(28.4%)	1.81(41.6%)	2.61(3.7%)	2.63(2.9%)	2.96(8.2%)	2.86(11.2%)	2.81(20.6%)	2.79(21.3%)
US Births	674.54(−2.1%)	681.81(−3.2%)	859.89(−2.8%)	866.45(−3.6%)	781.99(0.5%)	768.68(2.2%)	1417.01(−0.3%)	1459.69(−3.3%)

(b) MAE (Improvement, %)

Family	TiRex		LLMTime		LSTPrompt	
	85%	70%	85%	70%	85%	70%
Bitbrains-FS	48.45(83.2%)	10.06(96.5%)	424.21(13.5%)	3.49(99.3%)	429.39(17.3%)	329.79(36.5%)
Bitbrains-Rnd	24.94(35.9%)	16.58(57.4%)	18.92(30.3%)	4.56(83.2%)	60.01(56.8%)	49.63(64.3%)
Bizitobs-App	1492.75(−1.8%)	1296.77(11.6%)	888.52(8.1%)	694.13(28.2%)	1303.01(48.2%)	1252.66(50.2%)
Bizitobs-L2C	7.97(5.1%)	7.49(10.8%)	7.44(−1.8%)	6.99(4.4%)	8.38(11.0%)	8.39(10.9%)
Bizitobs-Svc	29.70(16.4%)	16.90(52.4%)	43.66(7.1%)	10.85(76.9%)	51.79(8.6%)	12.59(77.8%)
Electricity	346.86(9.6%)	256.13(33.2%)	951.81(0.2%)	950.87(0.3%)	1343.93(3.5%)	850.56(38.9%)
ETT1	20.24(13.6%)	15.29(34.7%)	40.41(22.6%)	31.43(39.8%)	57.13(13.2%)	52.56(20.1%)
ETT2	29.01(8.2%)	25.06(20.7%)	75.19(3.8%)	66.95(14.3%)	102.43(4.4%)	109.80(−2.5%)
Jena Weather	6.34(24.7%)	4.30(48.9%)	3.53(20.9%)	2.84(36.4%)	8.50(−2.2%)	8.00(3.9%)
KDD Cup 2018	9.01(0.6%)	9.14(−0.8%)	14.31(−2.0%)	14.92(−6.3%)	21.21(4.2%)	20.39(7.9%)
Loop Seattle	4.85(3.0%)	4.69(6.2%)	5.38(4.8%)	5.21(7.9%)	7.18(3.7%)	6.59(11.6%)
M4-Daily	118.51(14.3%)	88.03(36.3%)	161.85(13.5%)	125.39(32.9%)	270.35(11.2%)	241.97(20.5%)
M4-Hourly	234.46(18.5%)	173.13(39.8%)	655.08(15.6%)	449.05(42.1%)	803.54(49.9%)	844.95(47.3%)
M4-Monthly	730.67(−9.2%)	803.51(−20.1%)	707.74(9.9%)	801.68(−2.0%)	1870.51(3.5%)	2050.81(−5.8%)
M4-Quarterly	245.46(7.2%)	245.16(7.4%)	210.12(10.1%)	213.04(8.8%)	609.59(10.2%)	662.50(2.4%)
M4-Weekly	209.82(−1.7%)	129.07(37.4%)	225.56(26.0%)	184.77(39.4%)	204.26(22.7%)	169.61(35.8%)
M-Dense	60.08(9.5%)	59.72(10.0%)	136.89(10.8%)	124.12(19.1%)	222.87(5.8%)	194.89(17.6%)
Saugeen	14.23(1.9%)	13.00(10.4%)	18.04(3.3%)	19.06(−2.1%)	29.36(7.4%)	26.45(16.6%)
Solar	7.63(−1.9%)	7.96(−6.3%)	9.61(10.5%)	10.16(5.4%)	110.52(32.6%)	125.37(23.5%)
SZ-Taxi	2.51(−0.7%)	2.54(−1.7%)	2.63(3.5%)	2.75(−1.1%)	3.82(0.2%)	3.58(6.4%)
Temp-Rain	2.01(29.7%)	1.68(41.3%)	2.83(20.0%)	2.33(34.3%)	4.61(9.5%)	4.39(13.6%)
US Births	747.34(−0.2%)	742.44(0.5%)	1363.93(−5.3%)	1329.07(−2.6%)	3203.44(−6.6%)	3233.08(−7.6%)

Table 17: Pairwise-Point retained MAE at 85% and 70% target coverage by dataset family and backbone. Parentheses show percentage improvement over full-coverage family-mean MAE. TSFM results use the primary cohort; LLM-based results use the matched short-horizon cohort.

(a) MAE (Improvement, %)

Family	Chronos		TimeMoE		TimesFM		Moirai	
	85%	70%	85%	70%	85%	70%	85%	70%
Bitbrains-FS	352.86(−5.2%)	322.77(3.8%)	315.28(−2.1%)	386.67(−25.2%)	347.00(−3.7%)	427.00(−27.6%)	320.49(−3.7%)	299.10(3.2%)
Bitbrains-Rnd	37.25(−12.8%)	30.44(7.8%)	46.81(−12.7%)	43.30(−4.2%)	46.15(−12.6%)	40.22(1.9%)	49.34(−13.0%)	44.92(−2.9%)
Bizitobs-App	2050.26(−4.3%)	2209.91(−12.4%)	2477.55(9.9%)	2747.08(0.1%)	2824.77(6.7%)	3124.87(−3.3%)	2517.16(−0.6%)	2573.48(−2.9%)
Bizitobs-L2C	10.53(0.6%)	10.93(−3.2%)	10.21(−2.8%)	11.25(−13.3%)	10.46(1.7%)	10.25(3.7%)	11.41(2.9%)	11.54(1.8%)
Bizitobs-Svc	79.92(−3.2%)	84.78(−9.4%)	108.77(8.8%)	106.71(10.5%)	112.62(4.1%)	84.69(27.9%)	99.48(−0.5%)	85.81(13.3%)
Electricity	470.75(−1.7%)	439.16(5.2%)	449.73(−4.0%)	396.56(8.3%)	417.39(4.1%)	422.51(2.9%)	701.50(3.2%)	663.68(8.4%)
ETT1	25.17(6.9%)	26.22(3.0%)	23.89(0.3%)	21.98(8.3%)	23.51(0.4%)	21.66(8.3%)	27.16(0.7%)	24.83(9.2%)
ETT2	30.70(4.6%)	30.07(6.6%)	36.62(4.3%)	33.56(12.3%)	37.41(6.0%)	35.65(10.4%)	37.36(10.6%)	33.67(19.4%)
Jena Weather	6.02(42.9%)	5.32(49.5%)	6.11(46.4%)	4.84(57.5%)	6.27(38.0%)	5.29(47.7%)	8.78(48.7%)	8.13(52.5%)
KDD Cup 2018	9.92(12.7%)	9.95(12.4%)	9.52(−0.7%)	8.74(7.6%)	10.55(2.5%)	10.70(1.1%)	11.28(1.2%)	10.21(10.6%)
Loop Seattle	5.39(9.2%)	4.94(16.7%)	5.44(5.2%)	4.91(14.3%)	6.04(12.6%)	5.59(19.2%)	7.49(12.9%)	6.82(20.6%)
M4-Daily	130.92(23.9%)	117.76(31.5%)	172.44(18.6%)	167.00(21.2%)	152.66(11.8%)	140.76(18.7%)	144.38(12.6%)	132.13(20.0%)
M4-Hourly	349.26(3.3%)	131.56(63.6%)	213.61(29.4%)	208.33(31.1%)	203.01(41.2%)	190.03(45.0%)	888.57(7.9%)	551.54(42.8%)
M4-Monthly	472.60(30.3%)	455.86(32.7%)	611.48(46.8%)	576.94(49.8%)	453.66(2.9%)	477.84(−2.3%)	527.00(30.0%)	513.97(31.7%)
M4-Quarterly	205.91(28.4%)	128.65(55.3%)	178.71(50.0%)	186.32(47.9%)	236.53(5.3%)	173.58(30.5%)	188.25(29.4%)	183.49(31.2%)
M4-Weekly	198.64(3.9%)	79.24(61.7%)	229.90(65.1%)	204.61(68.9%)	209.03(8.7%)	167.81(26.7%)	231.91(14.6%)	203.62(25.0%)
M-Dense	72.02(−3.5%)	68.53(1.5%)	83.80(−2.1%)	83.59(−1.8%)	71.42(−5.1%)	72.55(−6.8%)	229.36(3.8%)	203.58(14.6%)
Saugeen	13.71(15.2%)	10.70(33.8%)	14.98(4.2%)	13.95(10.8%)	11.06(23.8%)	10.30(29.0%)	14.95(9.8%)	11.77(29.0%)
Solar	6.89(12.0%)	5.99(23.5%)	6.23(16.0%)	5.58(24.7%)	8.46(14.3%)	7.68(22.2%)	16.67(13.2%)	15.24(20.7%)
SZ-Taxi	2.64(0.8%)	2.80(−5.3%)	2.73(−5.0%)	2.87(−10.4%)	2.60(−2.6%)	2.70(−6.5%)	2.99(−1.1%)	3.02(−2.0%)
Temp-Rain	3.40(−9.6%)	2.91(6.2%)	2.72(−0.4%)	2.02(25.3%)	3.52(−9.2%)	3.11(3.5%)	3.75(−5.9%)	3.25(8.1%)
US Births	610.45(7.6%)	643.81(2.6%)	806.80(3.5%)	800.86(4.2%)	731.19(7.0%)	748.36(4.8%)	1380.96(2.2%)	1321.53(6.5%)

(b) MAE (Improvement, %)

Family	TiRex		LLMTime		LSTPrompt	
	85%	70%	85%	70%	85%	70%
Bitbrains-FS	296.88(−2.7%)	365.26(−26.3%)	399.58(18.5%)	487.70(0.5%)	587.38(−13.1%)	710.30(−36.8%)
Bitbrains-Rnd	43.91(−12.8%)	39.58(−1.6%)	30.35(−11.7%)	33.09(−21.8%)	123.68(11.0%)	136.17(2.0%)
Bizitobs-App	1374.98(6.3%)	1451.78(1.0%)	874.52(9.5%)	1025.80(−6.1%)	1342.01(46.6%)	1520.19(39.5%)
Bizitobs-L2C	8.41(−0.2%)	8.85(−5.4%)	7.59(−3.8%)	7.02(4.0%)	8.09(14.1%)	7.14(24.2%)
Bizitobs-Svc	36.27(−2.2%)	33.37(6.0%)	28.61(39.1%)	29.28(37.7%)	26.13(53.9%)	27.02(52.3%)
Electricity	383.34(0.0%)	378.96(1.2%)	965.88(−1.3%)	827.35(13.2%)	1361.16(2.3%)	1025.50(26.4%)
ETT1	23.81(−1.7%)	23.60(−0.8%)	55.60(−6.5%)	49.40(5.4%)	55.80(15.2%)	51.08(22.4%)
ETT2	29.97(5.1%)	28.33(10.3%)	61.78(20.9%)	61.70(21.0%)	97.83(8.7%)	72.54(32.3%)
Jena Weather	5.50(34.6%)	4.75(43.6%)	2.99(33.1%)	3.49(21.8%)	7.16(13.9%)	8.22(1.2%)
KDD Cup 2018	9.11(−0.5%)	8.48(6.5%)	11.48(18.1%)	10.43(25.6%)	14.51(34.4%)	14.40(35.0%)
Loop Seattle	4.70(6.2%)	4.30(14.1%)	4.73(16.4%)	4.26(24.7%)	5.77(22.7%)	5.41(27.5%)
M4-Daily	118.26(14.5%)	119.82(13.3%)	129.14(30.9%)	86.61(53.7%)	221.23(27.4%)	161.24(47.1%)
M4-Hourly	223.79(22.2%)	229.18(20.3%)	266.97(65.6%)	292.72(62.3%)	1148.97(28.4%)	1336.33(16.7%)
M4-Monthly	438.13(34.5%)	403.17(39.7%)	494.83(37.0%)	427.74(45.6%)	1203.23(37.9%)	841.19(56.6%)
M4-Quarterly	169.09(36.1%)	154.66(41.6%)	225.70(3.4%)	142.60(39.0%)	415.79(38.8%)	304.52(55.2%)
M4-Weekly	201.20(2.4%)	182.42(11.5%)	248.19(18.6%)	221.16(27.4%)	264.19(0.0%)	268.51(−1.6%)
M-Dense	69.48(−4.7%)	67.00(−1.0%)	161.44(−5.2%)	162.72(−6.0%)	231.51(2.1%)	234.93(0.7%)
Saugeen	11.29(22.1%)	11.08(23.6%)	16.34(12.5%)	12.73(31.8%)	24.55(22.6%)	22.45(29.2%)
Solar	6.34(15.3%)	5.90(21.2%)	9.68(9.8%)	8.68(19.1%)	24.26(85.2%)	21.18(87.1%)
SZ-Taxi	2.62(−4.9%)	2.76(−10.4%)	2.65(2.7%)	2.70(0.7%)	3.70(3.3%)	3.72(2.7%)
Temp-Rain	3.13(−9.1%)	2.57(10.4%)	3.90(−10.1%)	3.49(1.6%)	4.62(9.2%)	4.26(16.2%)
US Births	700.52(6.1%)	705.67(5.4%)	1161.53(10.3%)	1146.21(11.5%)	2525.21(16.0%)	2109.50(29.8%)

F HYPERPARAMETER SENSITIVITY AND COMPONENT ABLATIONS

We examine sensitivity to the direction-mismatch penalty, segment count, neutral-anchor pool, and number of textual evaluators, together with a risk-score permutation diagnostic. The experiments use stored outputs from the five primary TSFM backbones. Unless otherwise specified, evaluation covers 1,778 windows across 81 configurations and 22 families. Metrics and family-bootstrap confidence intervals follow Appendix C.

F.1 DIRECTION-PENALTY SENSITIVITY

We examine the direction-penalty parameter $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$ with $K = 3$ fixed. Because the original segment-level BGE text assignments were not retained, we re-encode the stored textual responses once and hold the resulting ordinal assignments fixed across all five settings. Only α varies within this analysis.

Table 18 reports U_{macro} values ranging from 0.420 to 0.423. At the default $\alpha = 0.75$, utility is 0.423, compared with 0.420 in the main evaluation. This sweep therefore measures sensitivity within the re-encoded representation rather than reproducing the original evaluation exactly. The small variation suggests limited sensitivity to α under these fixed assignments.

Table 18: Direction-penalty sensitivity with $K = 3$. Brackets indicate 95% family-bootstrap confidence intervals. The default is $\alpha = 0.75$.

α	U_{macro} [95% CI]	ρ_{macro}	I_{80} (%)
0.00	$0.423\ [0.274, 0.567]$	0.390	16.0
0.25	$0.422\ [0.273, 0.565]$	0.390	16.0
0.50	$0.420\ [0.271, 0.564]$	0.393	16.2
0.75	$0.423\ [0.273, 0.569]$	0.397	17.5
1.00	$0.423\ [0.272, 0.570]$	0.397	17.7

F.2 SEGMENT-COUNT SENSITIVITY

We evaluate $K \in \{1, 3, 5\}$ with $\alpha = 0.75$. Due to computational budget constraints, we use a two-agent ensemble comprising Gemini and GPT. Textual responses are mapped to ordinal levels using BGE, and the two agents' risk scores are averaged before ranking. Table 19 reports the results. Utility is similar at $K = 1$ and $K = 3$, with a higher point estimate at $K = 5$.

Table 19: Segment-count sensitivity for the Gemini+GPT ensemble at $\alpha = 0.75$. N denotes windows before replication across five backbones; F denotes evaluated families. Brackets indicate 95% family-bootstrap confidence intervals.

Agents	K	N	F	U_{macro} [95% CI]	ρ_{macro}	I_{80} (%)
Gemini+GPT	1	1,778	22	$0.351\ [0.210, 0.492]$	0.300	14.8
Gemini+GPT	3	1,778	22	$0.352\ [0.197, 0.509]$	0.333	14.6
Gemini+GPT	5	1,778	22	$0.372\ [0.219, 0.524]$	0.364	17.0

F.3 NEUTRAL-ANCHOR POOL SENSITIVITY

The BGE mapping uses 19 neutral-level anchors, compared with five anchors for each non-neutral level. To examine sensitivity to the neutral pool, we evaluate 50 randomly drawn subsets of five neutral anchors while retaining the non-neutral anchors.

Table 20 reports a mean utility of 0.392, with values ranging from 0.321 to 0.418. All tested subsets retain positive utility. The stored full-bank reference is 0.420. Because this reference uses stored scores while the reduced pools use re-encoded assignments, differences from the reference are descriptive and do not isolate the effect of anchor removal alone. The results show variation across neutral-anchor subsets, rather than invariance to anchor composition.

Table 20: Sensitivity to reducing the neutral-anchor pool from 19 to five anchors. The reduced-pool range is the minimum and maximum across 50 draws, not a confidence interval.

Condition	U_{macro}	I_{80} (%)
Full-bank reference	0.420	17.5
Reduced pool: mean	0.392	16.9
Reduced pool: minimum	0.321	—
Reduced pool: maximum	0.418	—

Table 21: Global risk-score permutation diagnostic using 999 permutations. The observed interval is a 95% family-bootstrap interval.

Quantity	Value
Observed U_{macro}	0.420 $[0.269, 0.567]$
Null mean U_{macro}	0.001
Null maximum U_{macro}	0.079
Permutations at or above observed	0/999
Monte Carlo p-value	0.001

F.4 Risk-Score Permutation Diagnostic

We test whether the observed alignment between risk and forecast error exceeds that obtained after random reassignment of the risk scores. For each of 999 permutations, the stored ensemble risk scores are shuffled globally across the 8,890 backbone–window instances, while errors and configuration–backbone memberships remain fixed. The selective utility is then recomputed.

The observed utility of 0.420 exceeds all permuted values. The null mean is approximately 0.001, and the maximum is 0.079. Using

$$p = \frac{1 + \#\{U^{\text{perm}} \geq U^{\text{obs}}\}}{999 + 1},$$

the Monte Carlo permutation p-value is 0.001.

This diagnostic supports the observed risk–error association relative to global random score reassignment. It does not specifically isolate the contribution of text–forecast pairing, because the permutation is applied to the completed risk scores rather than to the textual signals used to construct them.

F.5 Sensitivity to Agent Count

We evaluate all 15 nonempty subsets of the four Agent A models on the primary cohort. For each subset, risk is the equal-weight mean of its members' scores before ranking. Table 22 reports every subset, and Table 23 summarizes results by subset size. GPT, Gemini, Fable, and DeepSeek denote GPT-6 Astra, Gemini 3.6 Flash, Fable 5.1 Max, and DeepSeek V4 Pro, respectively.

Mean utility across subsets increases from 0.363 for one agent to 0.399, 0.415, and 0.419 for two, three, and four agents. However, agent composition also matters: some two- and three-agent subsets have higher utility point estimates than the four-agent ensemble. Thus, the observed group-size trend does not establish that adding an agent always improves performance or that four agents are optimal.

The four-agent result recomputed in this sweep is 0.419; the stored primary reference is 0.420. Both are reported separately. Ranges across subsets describe variation in agent composition and are not confidence intervals.

Table 22: Reliability results for all nonempty Agent A subsets on the primary cohort. Brackets indicate 95% family-bootstrap confidence intervals.

m	Agents	U_{macro} [95% CI]	ρ_{macro}	I_{80} (%)
1	Fable	$0.390\ [0.240, 0.540]$	0.369	15.0
1	DeepSeek	$0.388\ [0.267, 0.506]$	0.345	14.9
1	Gemini	$0.343\ [0.175, 0.505]$	0.327	14.4
1	GPT	$0.331\ [0.193, 0.467]$	0.312	12.9
2	DeepSeek+Fable	$0.432\ [0.292, 0.567]$	0.390	16.7
2	DeepSeek+Gemini	$0.414\ [0.268, 0.554]$	0.373	16.2
2	Fable+Gemini	$0.402\ [0.243, 0.559]$	0.383	15.7
2	DeepSeek+GPT	$0.401\ [0.265, 0.534]$	0.361	16.1
2	Fable+GPT	$0.392\ [0.244, 0.537]$	0.374	15.1
2	Gemini+GPT	$0.352\ [0.194, 0.512]$	0.333	14.6
3	DeepSeek+Fable+Gemini	$0.433\ [0.282, 0.579]$	0.405	17.2
3	DeepSeek+Fable+GPT	$0.423\ [0.281, 0.561]$	0.387	16.9
3	DeepSeek+Gemini+GPT	$0.408\ [0.262, 0.551]$	0.378	16.3
3	Fable+Gemini+GPT	$0.397\ [0.240, 0.552]$	0.379	15.5
4	All four: recomputed	$0.419\ [0.268, 0.566]$	0.396	17.5
4	Primary reference	$0.420\ [0.269, 0.567]$	0.396	17.5

Table 23: Agent-count summaries. Each mean weights subsets of the same size equally; ranges are minimum–maximum utility across those subsets.

m	Subsets	Mean U	Range U	Mean ρ	Mean I_{80} (%)
1	4	0.363	$[0.331, 0.390]$	0.338	14.3
2	6	0.399	$[0.352, 0.432]$	0.369	15.7
3	4	0.415	$[0.397, 0.433]$	0.387	16.5
4	1	0.419	$[0.419, 0.419]$	0.396	17.5

G Complete Case Studies and Empirical Failure Modes

G.1 Complete High- and Low-Consistency Examples

Figure 7 presents additional examples for Chronos, Moirai, TiRex, and LSTPrompt. Each backbone contributes one relatively high-consistency (low-risk) and one relatively low-consistency (high-risk) forecast. These examples are illustrative and were selected without a prespecified selection rule. Each panel shows the observed context, realised future, and point forecast, together with the single-agent C_{risk} and realised MAE.

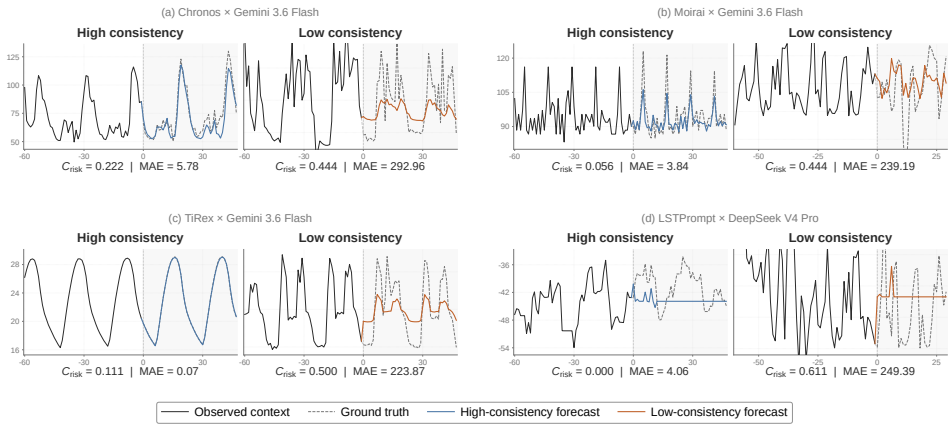


Figure 7: Additional high- and low-consistency forecasts for Chronos, Moirai, TiRex, and LSTPrompt. Scores are computed using the individual Agent A models listed in Table 24. The selected examples illustrate contrasting risk scores and realised errors; they do not establish the prevalence of risk–error alignment.

Table 24 lists the metadata for the eight examples. High and Low refer to consistency, so High corresponds to the lower risk score within each backbone's pair.

Table 24: Metadata for the examples in Figure 7. All examples use context length $T = 512$ and horizon $H = 48$. High and Low denote relative consistency within each pair.

Cond.	Backbone	Agent A	Model class	Dataset	Freq.	C_{risk}	MAE
High	Chronos	Gemini 3.6 Flash	TSFM	bizitobs_l2c	5 min	0.222	5.78
Low	Chronos	Gemini 3.6 Flash	TSFM	bizitobs_l2c	5 min	0.444	292.96
High	Moirai	Gemini 3.6 Flash	TSFM	bizitobs_l2c	hourly	0.056	3.84
Low	Moirai	Gemini 3.6 Flash	TSFM	bizitobs_l2c	hourly	0.444	239.19
High	TiRex	Gemini 3.6 Flash	TSFM	m4_hourly	hourly	0.111	0.067
Low	TiRex	Gemini 3.6 Flash	TSFM	m4_hourly	hourly	0.500	223.87
High	LSTPrompt	DeepSeek V4 Pro	LLM-based	bizitobs_l2c	5 min	0.000	4.06
Low	LSTPrompt	DeepSeek V4 Pro	LLM-based	bizitobs_l2c	5 min	0.611	249.39

G.2 False Reassurance

We quantify false reassurance as the high-error share within the lowest-risk 20% of windows, where high error denotes the highest-MAE 20% within each configuration–backbone group. Cutoff ties receive fractional inclusion weights so that each set has the specified mass. Under independent random retention, the expected rate is 0.20.

On the primary cohort of 8,890 backbone–window instances, using the hierarchical aggregation in Appendix C, CoRE-Ensemble achieves a false reassurance rate of 0.111 (95% family-bootstrap CI $[0.068, 0.158]$), compared with 0.116 for Pairwise-Point. The paired difference is -0.005 (95% CI $[-0.070, 0.072]$), providing no statistically resolved advantage over Pairwise-Point. Although CoRE's observed rate is below the Random reference, high-error forecasts remain present among its low-risk predictions.

H Reproducibility, Computational Resources, and Licenses

H.1 Software and Reproducibility

The evaluation pipeline uses Python 3.10 and PyTorch, together with the forecasting backbones' inference libraries. Model configurations and scoring details are described in Appendix B. Stored point forecasts and Agent A responses are reused for downstream scoring and evaluation.

The four Agent A models provide one stored response per primary-cohort window, giving 7,112 responses across 1,778 windows. These responses are reused across forecasting backbones, and the matched short-horizon analyses reuse the corresponding subset. This count describes the stored responses rather than the total number of API attempts, including retries. Reproducibility of the external API outputs relies on the preserved responses rather than deterministic regeneration.

H.2 Computational Resources

Experiments were conducted on an Ubuntu system equipped with two NVIDIA GeForce RTX 5090 GPUs with approximately 32 GB of memory each. GPU-based forecasting runs used one GPU at a time, with one evaluation window per inference call. TimesFM and the BGE text encoder ran on CPU. LLMTime and LSTPrompt used quantized LLaMA-2-13B models through `llama.cpp` and Ollama, respectively.

Recorded Agent A token usage totals 2,914,789 tokens: 463,964 for Gemini 3.6 Flash, 431,650 for GPT-6 Astra, 1,060,041 for Fable 5.1 Max, and 959,134 for DeepSeek V4 Pro. These totals reflect usage recorded in the stored response metadata and do not establish total usage across retries. Production runs were not instrumented for end-to-end latency or GPU-hour accounting. We therefore do not report runtime estimates or monetary costs.

H.3 Licenses and Terms of Use

Evaluation data were accessed through the GIFT-Eval collection and remain subject to the terms of their original sources. Explicit licenses could not be established for some dataset families; inclusion in the collection should not be interpreted as permission to redistribute them.

Forecasting backbones and the BGE encoder are subject to their respective model licenses. Agent A services are accessed under the corresponding API provider terms. Our repository does not redistribute third-party dataset contents or model weights.

Broader Impacts

CoRE-TSF supports selective forecasting but does not guarantee the accuracy of retained predictions. Its use in consequential applications requires application-specific validation and appropriate oversight.