What If TSF: REFRAMING TIME SERIES FORECASTING AS SCENARIO-GUIDED MULTIMODAL FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in large language models (LLMs) have enabled time series forecasting to move beyond numerical observations and incorporate external information in multimodal settings. Such information can improve forecasting performance, but accuracy alone may not reveal whether models appropriately respond to it: a model may ignore relevant information, fail to distinguish scenarios with different implications, or overreact to irrelevant or weak signals. We introduce *What If TSF (WIT)*, a benchmark for evaluating whether models effectively respond to future scenarios. WIT constructs *controlled scenario sets* by fixing the forecasting state while systematically varying future scenarios, enabling three complementary evaluations: *Factual Comparison*, which measures the predictive utility of factual scenario information; *Relational Comparison*, which evaluates whether forecasts satisfy expected direction, contrast, intensity, and restraint relations across scenarios; and *Grounded Comparison*, which assesses whether scenario-induced revision trajectories are empirically plausible relative to comparable real cases. Experiments show that factual accuracy gains do not consistently translate into appropriate responses to alternative scenarios, demonstrating the need to evaluate multimodal forecasting beyond predictive accuracy.

1 INTRODUCTION

Most forecasting methods infer future dynamics primarily from historical numerical observations, either by modeling dependencies across observed variables or by learning transferable temporal patterns at scale with time-series foundation models (TSFMs) (Liu et al., 2024c; Wang et al., 2025; Jang et al., 2026; Liu et al., 2025a;b; 2026b; Khwaja et al., 2026). Recently, covariate-aware TSFMs have extended this scope beyond historical observations by conditioning forecasts on known future covariates, such as calendar features and planned interventions available over the prediction horizon (Ansari et al., 2025; Jain &Sen, 2026; Podest et al., 2026; Meyer et al., 2026). While this substantially broadens the information available to numerical forecasters, many consequential future drivers are difficult to represent as numerical covariates or may simply not be captured by the available covariate set. Information about expected shifts in supply or demand, as well as forecasts of unusual weather conditions, may instead be available through reports or news articles even when the time series itself provides no indication of such external changes. Forecasting in such settings therefore calls for models that can complement time series with external textual context.

Recent advances in large language models (LLMs) and multimodal alignment have made it increasingly feasible to incorporate natural-language context into time series forecasting (Kong et al., 2025). A growing body of work explores prompting, representation fusion, and context-aware forecasting methods that allow external information to influence numerical predictions (Requeima et al., 2024; Jia et al., 2024; Jin et al., 2024; Liu et al., 2024b; Li et al., 2026; Wu et al., 2026; Das et al., 2026). Benchmarks such as Time-MMD (Liu et al., 2024a) and TimesX (Liu et al., 2026a) further support this direction by providing richer and better-aligned textual context. Together, these developments have established an important premise: external context can provide predictive information beyond numerical extrapolation and can potentially improve forecasting performance.

Yet, a model can predict the observed future more accurately without necessarily reflecting what the provided context implies. A factual scenario may reduce forecast error relative to a no-scenario baseline, establishing its predictive utility without showing whether the forecast reflects the implica-

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.

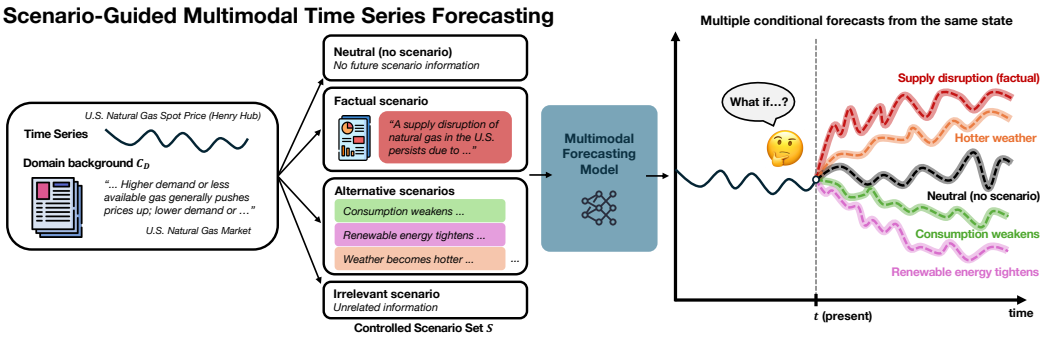


Figure 1: Formulation of scenario-guided multimodal time series forecasting. The figure highlights the role of scenario context in shaping forecasts. Forecasts are generated from controlled scenario sets with fixed histories and domain background context.

tions of that scenario. A model can therefore perform well on factual forecasting while still failing to distinguish how alternative contexts should change its forecasts. Thus, even when context is genuinely relevant and useful for prediction, an important question remains: *when only the future scenario changes, does the forecast revise in a way that is consistent with that scenario?*

Recent work has begun to move beyond factual context evaluation. CiK (Williams et al., 2025) evaluates forecasting with future information, constraints, and other types of context, while TADiff (Gu et al., 2026) explicitly studies counterfactual forecasting under alternative future conditions. However, these settings do not directly isolate scenario-induced forecast changes by holding the forecasting state fixed and comparing each scenario against the same neutral forecast. For example, a scenario with an upward implication should induce an upward revision, opposing scenarios should produce contrasting revisions, stronger versions of the same condition should produce larger revisions, and irrelevant information should leave the forecast largely unchanged. This motivates evaluating not only the predictive utility of context, but also whether forecasts respond consistently across scenarios and remain empirically grounded.

We introduce *What If TSF (WIT)* [1], a benchmark designed to evaluate whether forecasts reflect the implications of contextual information under controlled future scenarios. WIT evaluates forecasts using **controlled scenario sets**, holding the numerical history and domain background context fixed while systematically varying the future scenario (See Figure 1). By expanding evaluation from individual text-conditioned forecasts to structured comparisons under the same forecasting state, WIT assesses both factual predictive utility and consistency in scenario-induced forecast responses.

WIT evaluates these forecasts through three complementary comparisons. **Factual Comparison** asks how accurately the model predicts the observed future, how much factual scenario information reduces forecast error on average, and how often that information improves the prediction. **Relational Comparison** tests whether scenario-induced revisions relative to the model's neutral forecast satisfy explicit expectations about *Direction*, *Contrast*, *Intensity*, and *Restraint*. Finally, **Grounded Comparison** assesses how closely these neutral-relative revision trajectories align with deviations observed in comparable real cases. Together, these comparisons evaluate not only whether context improves observed-future prediction, but whether forecasts respond coherently and in empirically plausible ways as future scenarios change.

2 Related Work

Context-aware forecasting datasets and benchmarks. Time-MMD (Liu et al., 2024a) temporally aligns real-world time series with text collected from reports and web search results, evaluating text-conditioned forecasts against numerical-only baselines. CiK (Williams et al., 2025) manually constructs context across five information types—intemporal, future, historical, covariate, and causal—and evaluates context effects using Region of Interest (RoI)-weighted errors and constraint penalties, including removal and irrelevant or misleading context. Its tasks use both observed and synthetic numerical futures. GIFT-CTX (Yan et al., 2026) combines selected CiK examples with

[1] Data: https://anonymous.4open.science/r/WhatIfTSF-8E36/

Table 1: Comparison with prior context-aware forecasting datasets and benchmarks. **WIT evaluates scenario responses relative to a neutral forecast and assesses their empirical alignment with deviations observed in matched real cases.**

Dataset / Benchmark	Context comparisons	Context effect evaluation	Numerical reference
Time-MMD	Whole context	Error	Observed
CiK	Types + irrelevant or misleading	RoI* error + penalties	Observed / synthetic
GIFT-CTX	Whole context	RoI* error	Observed / synthetic
CAF-7M	Whole context + swapped	Error	Observed
TimesX	Types	Error	Observed
TADiff	Scenarios; factual / alternative	Error, consistency	Observed / synthetic
WIT	**Scenarios (each vs. neutral); factual / alternative / irrelevant**	Error, win rate†, **neutral-relative relations**	**Observed (Empirical matched sets)**

*Region of Interest; CiK additionally includes constraint penalties.

†Win-rate metrics in CiK and CAF-7M (DoubleCast) measure how often a model outperforms a forecasting baseline. WIT's Context Win Rate instead measures, for the same model and case, whether adding the factual scenario reduces forecast error relative to its neutral forecast.

additional synthetic cases designed so that accurate forecasting requires both numerical history and textual context. CAF-7M (Zheng et al., 2026) generates synthetic context from observed histories and futures, retains cases where context improves forecasting, and compares original, absent, and swapped context. TimesX (Liu et al., 2026a) constructs context from observed sources, organizing metadata, calendar, covariate, and event information and evaluating their incremental effects through forecast errors. TADiff (Gu et al., 2026) extends evaluation to factual and alternative scenarios under a shared numerical history. For real-world data, alternative conditions are selected from other samples, while its Synth dataset provides synthetic futures for alternative scenarios. Its evaluation combines forecast error with consistency scoring that measures the alignment of forecasts with intrinsic historical features and extrinsic future conditions.

Positioning WIT. Table 1 compares prior work in terms of context comparisons, context effect evaluation, context basis, and numerical references. Among different types of context, scenarios are especially important because they describe future conditions that may affect the forecast. Prior benchmarks show that context can improve forecast accuracy, but forecast error provides limited insight into how a model responds across scenarios, such as the direction, relative ordering, strength, and restraint of its revisions. Moreover, when an alternative scenario did not actually occur, its synthetic future must be constructed under additional assumptions, making the resulting trajectory difficult to justify or validate against real observations.

WIT therefore keeps numerical history and domain background context fixed, compares scenario-induced revisions against the same neutral forecast, and evaluates their relations using real-case-informed alternatives. These responses are further grounded by comparing their neutral-relative revision trajectories with deviations observed in matched real cases, without relying on synthetic alternative futures.

3 What If TSF Benchmark Construction

WIT follows a reusable construction protocol rather than relying on a fixed collection of manually curated forecasting cases. Given a target time series and a corpus of time-stamped textual sources, the protocol constructs the domain background context, forecasting anchors, factual scenarios, controlled scenario sets, and empirical matched sets through a common sequence of numerical and semantic checks. Although time series and textual context are domain-specific, all benchmark components are constructed under the same principles. This allows WIT to be extended to new targets and observation periods without redesigning the benchmark case by case.

Construction proceeds in four stages. *First*, we specify the forecasting target and relevant external mechanisms and generate candidate forecasting states. *Second*, frozen history-only forecasters identify candidates whose observed futures show persistent deviations from calibrated numerical

expectations. *Third*, source evidence is verified and condensed into factual scenarios, after which evidence-qualified candidates are retained as anchors. *Finally*, the same domain background is used to construct controlled scenario sets and empirical matched sets.

3.1 Domain Background Context and Candidate States

For each domain D, we define a domain background context $\mathcal{C}_D$ containing the target definition and a compact set of mechanisms through which external conditions can plausibly affect it. In the current release (Figure 2), the targets are the Henry Hub natural gas spot price (Energy), the Nominal Broad U.S. Dollar Index (Economy), and reported polling support for a selected party (Politics). Details of data sources and benchmark statistics are reported in Appendix A.

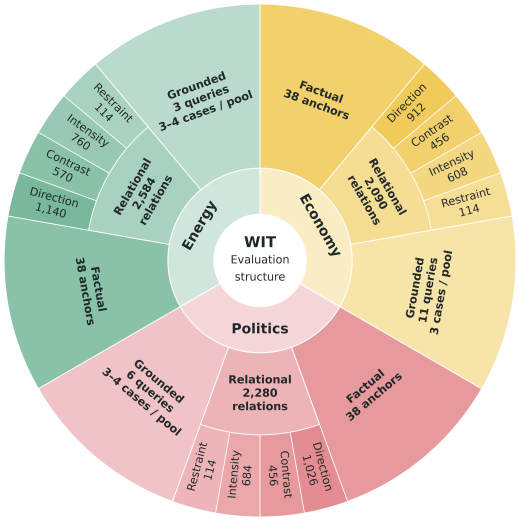


Figure 2: Current release of WIT benchmark composition by domain and evaluation category.

$\mathcal{C}_D$ serves both as the shared background supplied to models and as the semantic reference used during construction. It contains only general domain knowledge and no episode-specific future information. Each candidate state is $z_i = (\mathbf{x}_i, \mathcal{C}_D)$, where $\mathbf{x}_i$ is the numerical history. Candidate windows are generated from available observations while preserving the original observation calendar, and the same z_i is reused across all scenario conditions so that only the future scenario changes. Detailed window construction and domain-background generation are provided in Appendices B.1 and B.2, respectively. Examples of $\mathcal{C}_D$ for each domain are shown in Appendix C.2.1.

3.2 Reference-Guided Candidate Screening

Not every forecasting window is equally informative for evaluating scenario responsiveness. WIT therefore uses frozen history-only forecasts from Chronos-2, TimesFM-3, and TiRex-2 as construction references. We identify candidates with persistent same-direction deviations from the observed future across the three references.

To make these deviations comparable across domains, we robustly center and scale the reference-relative residuals at each forecast step. Because stride-one windows can share much of the same history and future, calibration uses only non-overlapping construction windows for each candidate, avoiding highly redundant examples. Cross-model agreement reduces dependence on any single forecaster, while the persistence criterion filters out cases whose deviation is concentrated in only a small part of the forecast horizon.

This stage yields numerically informative candidates, which become eligible for final anchor selection only after passing the factual evidence review. The construction-only calibration, cross-fitted reference-relative deviations, and candidate eligibility criteria are detailed in Appendix B.3.

3.3 Evidence-to-Scenario Construction

For each numerically eligible candidate, time-stamped textual sources covering the forecast horizon are examined for source-supported conditions relevant to mechanisms in $\mathcal{C}_D$. We separate source-grounded evidence extraction from scenario composition. Events and conditions are first extracted without exposing the observed target trajectory, reference-relative deviations, or any preferred response direction (Appendix B.4.2).

Factual scenarios are then composed from source-supported conditions according to source support, target relevance, temporal relevance, and mechanism applicability. Importantly, scenario composition does not use the observed reference-relative deviations. Any qualitative alignment between the

fixed scenario and those deviations is assessed only afterward and does not affect scenario wording or sample inclusion (Appendix B.4.3). This prevents factual scenarios from being selected simply because they align with the direction of the observed future.

The verified evidence is condensed into a *compact, mechanism-preserving scenario* containing only the information needed to characterize the relevant future condition. Observed target trajectories, reference forecasts, quantified target effects, and explicit forecasting instructions are excluded, while unnecessary identifying details are removed without discarding information required to interpret the scenario.

LLM assistance is used for evidence extraction and scenario composition, followed by a separate source and semantic review. Candidates without sufficiently supported factual scenarios are excluded from anchor selection. The final anchor set is then selected from evidence-qualified candidates to cover the observation period broadly while avoiding near-duplicate cases within the same forecasting entity (Appendix B.5). Detailed review and sample-acceptance criteria are provided in Appendix B.4.4.

3.4 Controlled Scenarios and Empirical References

Controlled scenario set. The controlled scenario set is constructed once from the fixed $\mathcal{C}_D$ and reused across anchors. Each scenario explicitly includes the information required to determine its qualitative implication, so that its relation can be interpreted without relying on additional anchor-specific context. An expected upward or downward revision defines *Direction*; paired scenarios with opposing effects define *Contrast*; stronger and weaker versions of the same condition define *Intensity*; and scenarios unrelated to the target define *Restraint*. Only relations that can be directly determined from the stated scenarios and $\mathcal{C}_D$ are retained (Appendix B.6). This enables controlled comparisons without constructing synthetic alternative futures.

Empirical matched sets. Grounded Comparison uses separately constructed matched sets of real cases. For each supported anchor–scenario pair, matched cases must have a similar numerical history, similar frozen construction-reference forecasts, and source evidence that the same scenario mechanism occurred with compatible timing, scope, roles, and applicability. Matching does not use the observed future. The matched cases therefore provide empirical references for neutral-relative revision trajectories, rather than counterfactual outcomes for the anchor or estimates of causal effects. Further matching details are provided in Appendix B.7.

4 Benchmark Evaluation Details

WIT evaluates scenario-guided forecasting through three complementary comparisons: *Factual*, *Relational*, and *Grounded* (see Figure 3). This separation reflects a fundamental asymmetry between factual and alternative scenarios. The factual scenario represents source-supported conditions that actually occurred or remained relevant during the forecast horizon and therefore has an observed future against which forecast accuracy can be measured. An alternative scenario did not occur for the same anchor, so constructing a synthetic future for it would require additional assumptions about the unobserved outcome. WIT instead evaluates alternative scenarios through neutral-relative relations that can be specified without constructing synthetic futures and, where empirical support is available, through comparable real cases in which the same mechanism occurred. Further evaluation details and a concrete Economy-domain example are provided in Appendices D and Figure 5.

4.1 Neutral-Relative Scenario Revisions

Let $\widehat{\mathbf{y}}_i^{\,s} = f_\theta(\mathbf{x}_i, \mathcal{C}_D, s)$ denote the forecast conditioned on scenario s, and let $\widehat{\mathbf{y}}_i^{\,0} = f_\theta(\mathbf{x}_i, \mathcal{C}_D, \varnothing)$ denote the corresponding *neutral forecast*. Neutral means scenario-unspecified; it does not represent an event-free counterfactual. We define the scenario-induced revision relative to the model's own neutral forecast as

$$\Delta_{i,h}(s) = \widehat{y}_{i,h}^{\,s} - \widehat{y}_{i,h}^{\,0}, \qquad h = 1, \ldots, H. \tag{1}$$

This neutral-relative formulation focuses evaluation on scenario-induced forecast revisions while reducing sensitivity to differences in baseline forecast levels across models.

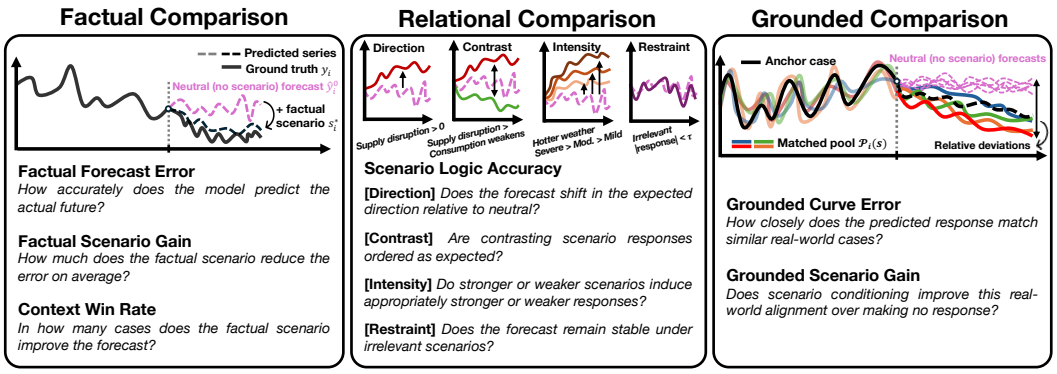


Figure 3: Three complementary comparisons in WIT benchmark.

4.2 Factual Comparison

Factual Comparison asks how accurately the model predicts the observed future $\mathbf{y}_i$ and whether providing the factual scenario improves that prediction relative to the neutral forecast. Let s_i^* denote the factual scenario for anchor i, and define

$$e_i^q = \frac{1}{H}\sum_{h=1}^{H}\left(\widehat{y}_{i,h}^q - y_{i,h}\right)^2, \qquad \overline{e}^{\,q} = \frac{1}{N}\sum_{i=1}^{N} e_i^q, \tag{2}$$

where $q = 0$ denotes the neutral condition and $q = s_i^*$ the factual condition.

We characterize factual performance from three complementary perspectives:

$$\mathrm{FFE} = \overline{e}^{\,s^*}, \quad \mathrm{FSG} = 100\left(1 - \frac{\overline{e}^{\,s^*}}{\overline{e}^{\,0}}\right), \quad \mathrm{CWR} = \frac{100}{N}\sum_{i=1}^{N}\mathbf{1}\left[e_i^{s_i^*} < e_i^0\right]. \tag{3}$$

Factual Forecast Error (FFE) measures forecast error when the factual scenario is provided. **Factual Scenario Gain (FSG)** measures the aggregate reduction in error relative to the same model's neutral forecast; positive FSG indicates lower mean forecast error under the factual scenario. FSG is computed from aggregate errors rather than by averaging anchor-specific percentage changes. **Context Win Rate (CWR)** measures how consistently the factual scenario reduces forecast error across anchors, with ties not counted as improvements. FFE, FSG, and CWR therefore capture factual forecast accuracy, aggregate improvement, and the share of anchors that improve, respectively.

4.3 Relational Comparison

Relational Comparison evaluates whether neutral-relative forecast revisions satisfy predefined relations among controlled scenarios under $\mathcal{C}_D$. We consider four complementary properties: *Direction*, *Contrast*, *Intensity*, and *Restraint*.

$$R_i(s) = \frac{1}{H}\sum_{h=1}^{H}\Delta_{i,h}(s), \qquad A_i(s) = \frac{1}{H}\sum_{h=1}^{H}|\Delta_{i,h}(s)|. \tag{4}$$

Here, $R_i(s)$ summarizes the signed average revision induced by scenario s, while $A_i(s)$ measures its average absolute magnitude.

Let $d_s \in \{-1, +1\}$ denote the expected direction of scenario s, let (s_+, s_-) denote a pair of opposing scenarios, and let (s_k, s_{k+1}) denote adjacent intensity levels with common direction d. Using domain-specific practical tolerances η_D and τ_D, we evaluate

$$\begin{aligned} \text{Direction} : d_s R_i(s) > \eta_D, \quad \text{Contrast} : R_i(s_+) - R_i(s_-) > \eta_D, \\ \text{Intensity} : d\,[R_i(s_{k+1}) - R_i(s_k)] > \eta_D, \quad \text{Restraint} : A_i(s_{\text{irr}}) \le \tau_D. \end{aligned} \tag{5}$$

The current-release values, reusable procedure for setting these tolerances from domain-specific target scales, and sensitivity analyses are detailed in Appendix D.3.2.

Each relation probes a different aspect of scenario response and is evaluated using **Scenario Logic Accuracy (SLA)**. **Direction** checks for a non-negligible revision in the expected direction. **Contrast** asks whether opposing scenarios produce distinguishable revisions. **Intensity** captures whether stronger conditions lead to correspondingly stronger revisions. **Restraint** measures whether irrelevant information leaves the forecast essentially unchanged. Together, the four relations assess directional responsiveness, contrast between opposing scenarios, sensitivity to scenario strength, and restraint under irrelevant information.

Let $\mathcal{R}_{i,c}$ denote the relations for anchor i in category c, and let $\mathcal{A}_c$ denote the anchors eligible for that category. Relations are first averaged within each anchor, anchors are weighted equally within each category, and the four categories are weighted equally in the Overall SLA.

$$\mathrm{SLA}_c = \frac{100}{|\mathcal{A}_c|} \sum_{i \in \mathcal{A}_c} \frac{1}{|\mathcal{R}_{i,c}|} \sum_{r \in \mathcal{R}_{i,c}} \mathbf{1}[r], \qquad \mathrm{SLA} = \frac{1}{4} \sum_{c \in \{\mathrm{Dir,Con,Int,Res}\}} \mathrm{SLA}_c. \tag{6}$$

We report Direction, Contrast, Intensity, Restraint, and Overall separately so that different failure modes remain visible. SLA therefore measures consistency with predefined scenario relations without requiring a target trajectory for an unobserved alternative scenario.

4.4 Grounded Comparison

Relational Comparison establishes whether forecast revisions have the expected qualitative structure, but satisfying those relations does not ensure agreement with responses observed in comparable real cases. For example, a model may correctly order two scenarios while producing revisions whose magnitudes differ substantially from empirical references. Grounded Comparison therefore evaluates scenario-induced revisions against matched real cases.

For each supported anchor–scenario pair (i, s), let $\mathcal{P}_i(s)$ denote its empirical matched set, constructed as described in Section 3.4. For each matched case j, define

$$d_{j,h} = y_{j,h} - \widehat{y}^{0}_{j,h}, \tag{7}$$

where $\widehat{y}^{0}_{j,h}$ is the evaluated model's own neutral forecast for matched case j. Thus, $d_{j,h}$ represents the observed neutral-relative deviation in that matched case.

The pair-level mean squared curve discrepancy is

$$g_i(s) = \frac{1}{H|\mathcal{P}_i(s)|} \sum_{j \in \mathcal{P}_i(s)} \sum_{h=1}^{H} \left(\Delta_{i,h}(s) - d_{j,h}\right)^2. \tag{8}$$

The corresponding no-revision baseline sets $\Delta_{i,h}(s) = 0$:

$$g^0_i(s) = \frac{1}{H|\mathcal{P}_i(s)|} \sum_{j \in \mathcal{P}_i(s)} \sum_{h=1}^{H} (d_{j,h})^2. \tag{9}$$

Let $\mathcal{Q}_D$ denote the supported anchor–scenario pairs in domain D. We first average pair-level discrepancies within each domain:

$$\overline{g}_D = \frac{1}{|\mathcal{Q}_D|} \sum_{(i,s) \in \mathcal{Q}_D} g_i(s), \qquad \overline{g}^0_D = \frac{1}{|\mathcal{Q}_D|} \sum_{(i,s) \in \mathcal{Q}_D} g^0_i(s). \tag{10}$$

We report domain-wise **Grounded Curve Error (GCE)** and **Grounded Scenario Gain (GSG)**

$$\mathrm{GCE}_D = \sqrt{\overline{g}_D}, \qquad \mathrm{GSG}_D = 100 \left(1 - \frac{\overline{g}_D}{\overline{g}^0_D}\right). \tag{11}$$

Overall GSG is reported as the equal-domain macro-average of GSG_D.

GCE measures the discrepancy between the predicted scenario-induced revision and neutral-relative deviations observed in matched real cases across the forecast horizon. Because the alternative future is unobserved, matched cases serve as empirical references rather than ground truth: lower GCE indicates closer empirical alignment, but zero does not represent a uniquely correct alternative future. GSG measures the improvement in this alignment relative to making no revision; positive GSG indicates better agreement than the no-revision baseline.

Table 2: Factual Comparison on WIT. Factual Forecast Error (neutral forecast error) is reported separately for each domain. FSG and CWR are macro-averaged across the three domains.

Model	**Factual Forecast Error** (neutral forecast error)			**Factual Scenario Gain (%)**	**Context Win Rate (%)**
	Energy	**Economy**	**Politics**		
Directly prompted LLMs					
Mistral (7B)	$0.312_{\pm 0.013}$ (0.337 ± 0.012)	$2.596_{\pm 0.110}$ (2.805 ± 0.099)	$25.704_{\pm 0.924}$ (27.552 ± 0.840)	$7.19_{\pm 0.63}$	$55.56_{\pm 1.83}$
Llama3 (70B)	$0.276_{\pm 0.009}$ (0.313 ± 0.008)	$2.184_{\pm 0.072}$ (2.470 ± 0.066)	$22.792_{\pm 0.686}$ (25.424 ± 0.616)	$11.25_{\pm 0.69}$	$60.82_{\pm 2.21}$
GPT-4o	$0.254_{\pm 0.007}$ (0.298 ± 0.006)	$2.233_{\pm 0.061}$ (2.580 ± 0.061)	$19.614_{\pm 0.518}$ (23.500 ± 0.500)	$\mathbf{14.92_{\pm 0.64}}$	$65.50_{\pm 1.83}$
Context-aware forecasting systems					
DoubleCast	$0.245_{\pm 0.003}$ (0.274 ± 0.003)	$\mathbf{2.045_{\pm 0.021}}$ (2.300 ± 0.021)	$19.200_{\pm 0.210}$ (23.550 ± 0.240)	$13.38_{\pm 0.42}$	$66.08_{\pm 1.34}$
MoiraiAgent	$\mathbf{0.242_{\pm 0.007}}$ (0.272 ± 0.006)	$2.070_{\pm 0.058}$ (2.290 ± 0.055)	$\mathbf{18.950_{\pm 0.500}}$ (23.450 ± 0.490)	$13.28_{\pm 0.79}$	$\mathbf{67.25_{\pm 2.21}}$

5 Experiments and Findings

We evaluate WIT across three domains with distinct targets and relevant external mechanisms: Energy, Economy, and Politics. The current release contains 114 forecasting anchors, with 38 anchors per domain, constructed following the common protocol in Section 3. Each anchor consists of a 24-step numerical history and a 20-step forecast horizon, and is paired with the fixed domain background context, its factual scenario, and the domain's controlled scenario set. Supported anchor–scenario pairs are additionally linked to empirical matched sets for Grounded Comparison.

Following the evaluation framework in Section 4, our experiments examine whether gains under factual scenarios are accompanied by responses that satisfy predefined scenario relations and align with comparable real cases. We evaluate this through Factual, Relational, and Grounded Comparisons, testing whether factual forecasting improvement alone is sufficient to characterize scenario-guided forecasting behavior.

Models. We primarily compare two classes of scenario-conditioned forecasters. *Directly prompted LLMs*, including Mistral-7B-Instruct v0.3 (Jiang et al., 2023), Llama-3-70B-Instruct (Grattafiori et al., 2024), and GPT-4o (OpenAI, 2024), receive the numerical history, domain background context, and scenario through a standardized prompt and directly generate the forecast trajectory. This group tests whether general-purpose language models can produce scenario-dependent forecasts from numerical and textual inputs without a dedicated forecasting architecture.

We additionally evaluate two *context-aware forecasting systems*: DoubleCast (Zheng et al., 2026), using the cutoff-controlled Qwen1.5-14B-Chat text encoder and CAF-7M cross-modal alignment, and MoiraiAgent (Yan et al., 2026), using GPT-4o with Chronos-2 as its time-series forecasting tool. These systems explicitly couple textual context with time-series forecasting, providing specialized counterparts to direct LLM prompting. Exact model components, knowledge cutoffs, alignment details, and inference settings are provided in Appendix E.

Evaluation protocol. All scenario-conditioned models are evaluated on the same fixed anchors, scenarios, and empirical matched sets. Scale-dependent metrics are reported separately for Energy, Economy, and Politics, while dimensionless metrics are computed within each domain and macro-averaged with equal domain weight. Each configuration is evaluated over three runs, and we report the mean and sample standard deviation. Further details are provided in Appendices D and E.4.

Cross-comparison findings. The results reveal distinct patterns across the three comparisons. Full domain-wise results and additional analyses are provided in are provided in Appendix F. In Factual Comparison (Table 2), factual scenarios improve average forecast accuracy for all models, but not consistently across anchors. CWR shows that aggregate gains can coexist with many cases in which the factual scenario does not improve the forecast, distinguishing the magnitude of average improvement from the share of anchors that actually improve.

Table 3: Relational Comparison on WIT, macro-averaged across three domains.

Model	Scenario Logic Accuracy (%)				
	Direction	Contrast	Intensity	Restraint	**Overall**
Directly prompted LLMs					
Mistral (7B)	71.80±1.55	68.10±1.45	54.30±2.09	70.66±2.19	66.21±1.63
Llama-3 (70B)	80.90±1.05	77.60±1.10	65.21±1.65	79.53±1.75	75.81±1.23
GPT-4o	87.79±0.75	**84.61±0.85**	73.90±1.36	**85.09±1.46**	**82.85±0.98**
Context-aware forecasting systems					
DoubleCast	89.40±0.70	82.10±0.81	67.80±1.21	66.47±1.61	76.44±1.08
MoiraiAgent	**90.11±0.90**	83.90±0.99	**76.51±1.36**	71.83±1.61	80.59±1.11

Table 4: Grounded Comparison on WIT. Grounded Curve Error is reported separately for each domain, while Grounded Scenario Gain is macro-averaged across three domains.

Model	Grounded Curve Error			Grounded Scenario Gain (%)
	Energy	Economy	Politics	
Directly prompted LLMs				
Mistral (7B)	0.592 ±0.009	1.740 ±0.028	6.885 ±0.150	−4.56 ±1.17
Llama-3 (70B)	0.515 ±0.005	1.527 ±0.019	5.949 ±0.112	8.20 ±0.50
GPT-4o	0.491 ±0.009	1.449 ±0.023	5.538 ±0.094	11.38 ±0.86
Context-aware forecasting systems				
DoubleCast	**0.446 ±0.004**	**1.311 ±0.011**	4.902 ±0.050	**17.64 ±0.43**
MoiraiAgent	0.449 ±0.006	1.320 ±0.018	**4.827 ±0.078**	17.50 ±0.64

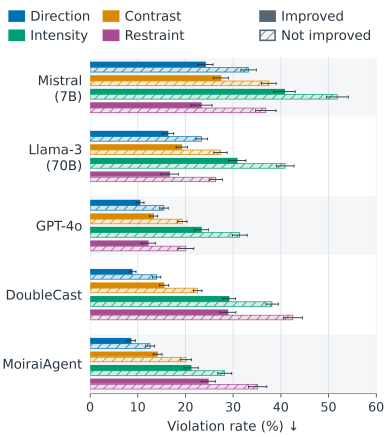


Figure 4: Relational violation rates for improved and not-improved anchors across three domains.

Relational Comparison (Table 3) indicates that appropriate responses to relevant scenarios do not necessarily imply restraint under irrelevant information or sensitivity to scenario strength. Similar Overall SLA can therefore hide important differences across response types.

Grounded Comparison (Table 4) shows a different pattern from Relational Comparison. The context-aware systems consistently achieve lower GCE across all three domains and higher GSG than the directly prompted LLMs, despite not achieving the highest Overall SLA. Thus, stronger relational consistency does not necessarily imply closer alignment with neutral-relative deviations observed in matched real cases.

Anchor-level analysis. We further examine whether factual improvement is associated with more consistent responses across scenarios (Figure 4; Appendix F). Anchors whose factual scenario improves forecast error generally show lower relational violation rates than those without improvement, although violations remain in both groups. Thus, factual improvement is associated with stronger relational consistency but still does not guarantee appropriate responses across all scenarios.

6 Conclusion

We introduced What If TSF (WIT), a benchmark for evaluating scenario-guided forecasting beyond accuracy on the observed future. WIT separates *Factual*, *Relational*, and *Grounded* comparisons to assess predictive gains, consistency with predefined relations across controlled scenario changes, and empirical alignment with matched real cases. Across three domains, factual gains do not ensure appropriate responses to alternative scenarios, and stronger relational consistency does not necessarily imply closer empirical alignment. These findings highlight the need to evaluate scenario responsiveness directly, rather than relying only on accuracy for the observed future.

AI USE STATEMENT

Generative AI tools were used for language polishing and stylistic refinement of the manuscript, and for selected benchmark-construction tasks, including source-grounded evidence extraction and factual scenario composition. AI-generated content was used only as intermediate assistance and not as a source of factual evidence. The authors take full responsibility for the final benchmark artifacts, manuscript content, and reported results.

ETHICS STATEMENT

WIT is intended as a research benchmark for evaluating scenario-guided forecasting and not as a decision-making system for financial, political, or policy applications. Source coverage and scenario construction may introduce biases, so benchmark results should be interpreted as comparative research evaluations rather than evidence for real-world decisions or policy actions.

REPRODUCIBILITY STATEMENT

Section 3 and the corresponding appendix describe the data sources and complete benchmark-construction protocol, including candidate screening, evidence-to-scenario construction, controlled scenarios, and empirical matching. Section 4 and Appendix D define the evaluation metrics, aggregation rules, and sensitivity analyses. Section 5 and Appendix E provide the evaluated model configurations, knowledge cutoffs, inference settings, prompts, and repeated-run protocol. Together, these specifications provide the information required to reproduce WIT construction and evaluation.

REFERENCES

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, et al. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.

Sarkar Snigdha Sarathi Das, Palash Goyal, Mihir Parmar, Nanyun Peng, Vishy Tirumalashetty, Chun-Liang Li, Rui Zhang, Jinsung Yoon, and Tomas Pfister. Nexus: An agentic framework for time series forecasting. *arXiv preprint arXiv:2605.14389*, 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Shuqi Gu, Yongxiang Zhao, Baoyu Jing, and Kan Ren. What if tomorrow is the world cup final? counterfactual time series forecasting with textual conditions. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=B8bAYc4COx.

Ayush Jain and Rajat Sen. Timesfm-3: A zero-shot foundation model for multivariate forecasting. Google Research Blog, 2026.

Jinkwan Jang, Hyungjin Park, Jinmyeong Choi, and Taesup Kim. Towards robust real-world multivariate time series forecasting: a unified framework for dependency, asynchrony, and missingness. In *International Conference on Learning Representations*, volume 2026, pp. 153495–153528, 2026.

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, pp. 23343–23351, 2024.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. Time-LLM: Time series forecasting by reprogramming large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Unb5CVPtae.

Emaad Khwaja, Chris Lettieri, Gerald Woo, Eden Belouadah, Marc Cenac, Guillaume Jarry, Enguerrand Paquin, Xunyi Zhao, Viktoriya Zhukov, Othmane Abou-Amal, et al. Toto 2.0: Time series forecasting enters the scaling era. *arXiv preprint arXiv:2605.20119*, 2026.

Yaxuan Kong, Yiyuan Yang, Shiyu Wang, Chenghao Liu, Yuxuan Liang, Ming Jin, Stefan Zohren, Dan Pei, Yan Liu, and Qingsong Wen. Achieving time series reasoning requires rethinking model design, tasks formulation, and evaluation. *arXiv preprint arXiv:2502.01477*, 2025.

Zihao Li, Xiao Lin, Zhining Liu, Jiaru Zou, Ziwei Wu, Lecheng Zheng, Dongqi Fu, Yada Zhu, Hendrik Hamann, Hanghang Tong, and Jingrui He. Language in the flow of time: Time-series-paired texts weaved into a unified temporal narrative. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=a1zBg9cBvt.

Chenghao Liu, Taha Aksu, Juncheng Liu, Xu Liu, Hanshu Yan, Quang Pham, Doyen Sahoo, Caiming Xiong, Silvio Savarese, and Junnan Li. Moirai 2.0: When less is more for time series forecasting. *arXiv preprint arXiv:2511.11698*, 2025a.

Haoxin Liu, Shangqing Xu, Zhiyuan Zhao, Lingkai Kong, Harshavardhan Prabhakar Kamarthi, Aditya Sasanur, Megha Sharma, Jiaming Cui, Qingsong Wen, Chao Zhang, et al. Time-mmd: Multi-domain multimodal dataset for time series analysis. *Advances in Neural Information Processing Systems*, 37:77888–77933, 2024a.

Haoxin Liu, Yichen Zhou, Rajat Sen, B. Aditya Prakash, and Abhimanyu Das. Rethinking multimodal time-series forecasting evaluation. In *Forty-third International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=auG2Cr76ed.

Xu Liu, Junfeng Hu, Yuan Li, Shizhe Diao, Yuxuan Liang, Bryan Hooi, and Roger Zimmermann. Unitime: A language-empowered unified model for cross-domain time series forecasting. In *Proceedings of the ACM Web Conference 2024*, pp. 4095–4106, 2024b.

Yiding Liu, Yifan Hu, Hongjie Xia, Peiyuan Liu, Hongzhou Chen, Xilin Dai, Zewei Dong, and Jiang-Ming Yang. Falcon-x: A time series foundation model for heterogeneous multivariate modeling. *arXiv preprint arXiv:2605.27286*, 2026b.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *International conference on learning representations*, volume 2024, pp. 11116–11140, 2024c.

Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models. In *Forty-second International Conference on Machine Learning*, 2025b. URL https://openreview.net/forum?id=LO7ciRpjI5.

Lucas Meyer, Claudio Sole, Huikan Xiang, Nicolas Li, Lucas Franceschino, Arnau Quera-Bofarull, Maarten P. Scholl, Joachim Fainberg, and Geoffrey Négiar. t_0: A time-series foundation model for forecasting with context. *arXiv preprint arXiv:2609.24559*, 2026.

OpenAI. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Patrick Podest, Marco Pichler, Elias Bürger, Levente Zólyomi, Bernhard Voggenberger, Wilhelm Berghammer, Daniel Klotz, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. Tirex-2: Generalizing tirex to multivariate data and streaming. *arXiv preprint arXiv:2607.01204*, 2026.

James Requeima, John Bronskill, Dami Choi, Richard Turner, and David K Duvenaud. Llm processes: Numerical predictive distributions conditioned on natural language. *Advances in Neural Information Processing Systems*, 37:109609–109671, 2024.

Shiyu Wang, Jiawei Li, Xiaoming Shi, Zhou Ye, Baichuan Mo, Wenze Lin, Shengtong Ju, Zhixuan Chu, and Ming Jin. Timemixer++: A general time series pattern machine for universal predictive analysis. In *International Conference on Learning Representations*, volume 2025, pp. 16980–17016, 2025.

Andrew Robert Williams, Arjun Ashok, Étienne Marcotte, Valentina Zantedeschi, Jithendaraa Subramanian, Roland Riachi, James Requeima, Alexandre Lacoste, Irina Rish, Nicolas Chapados, et al. Context is key: A benchmark for forecasting with essential textual information. In *Forty-second International Conference on Machine Learning*, 2025.

Xingjian Wu, Jianxin Jin, Wanghui Qiu, Peng Chen, Yang Shu, Bin Yang, and Chenjuan Guo. Aurora: Towards universal generative multimodal time series forecasting. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=VVJ6Ck9JBl.

Hanshu Yan, Hong-Quang Pham, Ibrahim Taha Aksu, Jun Hao Liew, Caiming Xiong, Silvio Savarese, Doyen Sahoo, and Junnan Li. Moiraiagent: An agentic framework for context-aware time-series forecasting. Salesforce AI Research Blog, 2026.

Vincent Zhihao Zheng, Étienne Marcotte, Arjun Ashok, Andrew Robert Williams, Lijun Sun, Alexandre Drouin, and Valentina Zantedeschi. Overcoming the modality gap in context-aided forecasting. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=NOHlqCnYot.

Table of Contents

A Data Sources and Benchmark Statistics **15**

A.1 Domain Data Sources . . . 15

A.1.1 Energy . . . 15

A.1.2 Economy . . . 15

A.1.3 Politics . . . 15

A.2 Benchmark Statistics . . . 16

A.2.1 Temporal Coverage and Window Statistics . . . 16

A.2.2 Anchor and Scenario Composition . . . 17

A.2.3 Controlled Scenario and Relation Inventory . . . 17

A.2.4 Empirical Matched-Set Coverage . . . 17

B Benchmark Construction Protocol **18**

B.1 Candidate Windows and Forecasting States . . . 18

B.2 Domain Background Context Construction . . . 18

B.3 Reference-Guided Candidate Screening . . . 19

B.3.1 Construction-Only Calibration Pool . . . 19

B.3.2 Cross-Fitted Reference-Relative Deviations . . . 19

B.3.3 Candidate Eligibility . . . 20

B.4 Factual Evidence-to-Scenario Construction . . . 20

B.4.1 Source-Supported Factual Evidence . . . 20

B.4.2 Outcome-Blind Event Extraction . . . 21

B.4.3 Outcome-Blind Factual Scenario Composition . . . 21

B.4.4 Source and Semantic Review . . . 21

B.4.5 Sample Acceptance . . . 21

B.5 Final Anchor Selection . . . 21

B.6 Controlled Scenario Set Construction . . . 22

B.7 Empirical Matched-Set Construction . . . 22

B.8 Leakage Controls, Freezing, and Audit Trail . . . 23

C Construction Prompts and Model Inputs **23**

C.1 Construction Prompt Templates . . . 23

C.1.1 Domain Background Context Construction . . . 23

C.1.2 Event Extraction . . . 23

C.1.3 Factual Scenario Composition . . . 24

C.1.4 Source and Semantic Review . . . 24

C.2 Model-Facing Input Examples . . . 24

C.2.1 Domain Background Context . . . 24

C.2.2 Factual Scenario . . . 24

C.2.3 Controlled Alternative Scenarios . . . 24

C.2.4 Irrelevant Scenario . . . 24

C.3 Evaluation Prompt Templates . . . 25

C.3.1 Directly Prompted LLMs . . . 25

C.3.2 Context-aware Forecasting Systems . . . 25

C.3.3 WIT Evaluation Example . . . 30

D Additional Evaluation Details **30**

D.1 Evaluation Populations and Aggregation . . . 30

D.2 Factual Comparison Details . . . 30

D.3 Relational Comparison Details . . . 31

D.3.1 Relation Eligibility and Scoring . . . 31

D.3.2 Practical Tolerances 31
D.3.3 Tolerance Sensitivity 32
D.4 Grounded Comparison Details 33
D.4.1 Empirical References for Evaluation 33
D.4.2 Pair and Matched-Case Weighting 33
D.4.3 No-Revision Baseline 33
D.4.4 Interpretation of GCE and GSG 33

E Experimental Setup **33**
E.1 Evaluated Models 33
E.2 Model Versions and Knowledge Cutoffs 33
E.3 Inference and Implementation Settings 34
E.4 Repeated-Run Protocol 35

F Additional Results and Analyses **35**
F.1 Numerical-Only Forecasting Baselines 35
F.2 Domain-Wise Factual Comparison 35
F.3 Domain-Wise Relational Comparison 35
F.4 Domain-Wise Grounded Comparison 36
F.5 Anchor-Level Analysis 36

G Limitations and Future Work **36**

A Data Sources and Benchmark Statistics

A.1 Domain Data Sources

A.1.1 Energy

WIT uses the *Henry Hub natural gas spot price* as the forecasting target for the *Energy* domain. We combine the numerical price series with textual information from official energy reports and news articles describing market conditions and relevant external events. The data collection period spans January 2, 2024 to September 9, 2026.

Time-series data. The numerical data consist of the Henry Hub Natural Gas Spot Price (NG.RNGWHHD.D), obtained from the U.S. Energy Information Administration (EIA) through its Open Data API [2]. The original observation calendar is preserved during benchmark construction; window construction and treatment of unavailable observations are described in Appendix B.1.

Textual sources. Textual information is collected from official EIA reports and energy news articles indexed through the Global Database of Events, Language, and Tone (GDELT).

- **Daily reports:** U.S. EIA, *Today in Energy*, tagged with Natural Gas, from January 4, 2024 to September 9, 2026.
  `https://www.eia.gov/todayinenergy/index.php?tg=natural%20gas`
- **Weekly reports:** U.S. EIA, *Natural Gas Weekly Update* and *Weekly Natural Gas Storage Report Supplement*, from January 11, 2024 to September 9, 2026.
  `https://www.eia.gov/naturalgas/weekly/`
  `https://www.eia.gov/naturalgas/weekly/supplement/`
- **News articles:** articles indexed through the GDELT 1.0 Global Knowledge Graph (GKG) from January 2, 2024 to September 9, 2026, retrieved using the keywords *natural gas* and *Henry Hub*.
  `http://data.gdeltproject.org/gkg/index.html`

A.1.2 Economy

WIT uses the *Nominal Broad U.S. Dollar Index* as the forecasting target for the *Economy* domain. We combine the numerical index series with textual information from news articles describing exchange rates, dollar conditions, and related macroeconomic developments. The data collection period spans January 2, 2024 to September 9, 2026.

Time-series data. The numerical data consist of the Nominal Broad U.S. Dollar Index (DTWEXBGS), obtained from the Federal Reserve Bank of St. Louis (FRED) [3]. The original observation calendar is preserved during benchmark construction; window construction and treatment of unavailable observations are described in Appendix B.1.

Textual sources. Textual information is collected from domestic and international news articles indexed through the Global Database of Events, Language, and Tone (GDELT).

- **News articles:** articles indexed through the GDELT 1.0 Global Knowledge Graph (GKG) from January 2, 2024 to September 9, 2026, retrieved using the keywords *dollar index*, *USD index*, and *exchange rate*.
  `http://data.gdeltproject.org/gkg/index.html`

A.1.3 Politics

WIT uses reported polling support for selected political parties as the forecasting target for the *Politics* domain. The numerical data cover multiple party-specific series across several countries

[2] `https://www.eia.gov/opendata/`
[3] `https://fred.stlouisfed.org/series/DTWEXBGS`

and are observed at varying intervals. We combine these series with country-specific news articles describing political and related external conditions relevant to party support. The data collection period spans January 2, 2024 to August 21, 2026.

Time-series data. The numerical data consist of reported polling support for major political parties, compiled by Europe Elects[4]. For each country, we select two to three parties with substantial polling support during the collection period. The selected parties are:

- **United Kingdom:** Conservative Party (CON), Labour Party (LAB), Reform UK (REFORM).
- **Hungary:** FIDESZ–Magyar Polgári Szövetség–Kereszténydemokrata Néppárt (Fidesz+KDNP), Tisztelet és Szabadság Párt (TISZA).
- **Germany:** Alternative für Deutschland (AfD), Christlich Demokratische Union Deutschlands–Christlich-Soziale Union in Bayern e.V. (CDU+CSU).
- **Greece:** Νέα Δημοκρατία (ND), ΠΑΣΟΚ - Κίνημα Αλλαγής (PASOK-KINAL), Συνασπισμός Ριζοσπαστικής Αριστεράς – Προοδευτική Συμμαχία (SYRIZA).

The original observation calendar of each party-specific series is preserved during benchmark construction; window construction for irregularly observed series is described in Appendix B.1.

Textual sources. Textual information is collected from major domestic and international news outlets for each country. Articles are retrieved according to country and target relevance, with detailed evidence filtering and factual-scenario construction described in Appendix B.4.1. The primary news sources are:

- **United Kingdom:** The Guardian[5], BBC[6], and The Telegraph[7].
- **Hungary:** Magyar Nemzet[8], Népszava[9], and 24.hu[10].
- **Germany:** Der Spiegel[11], Die Zeit[12], and Frankfurter Allgemeine Zeitung[13].
- **Greece:** Kathimerini[14], To Vima[15], and Ta Nea[16].

A.2 Benchmark Statistics

The current WIT release contains three domains—Energy, Economy, and Politics—with 38 forecasting anchors per domain and 114 anchors in total. All domains use a 24-step numerical history and a 20-step forecast horizon. The forecasting targets are the Henry Hub natural gas spot price in the Energy domain, the Nominal Broad U.S. Dollar Index in the Economy domain, and reported polling support for a selected party in the Politics domain. The scenario inventory differs by domain according to the mechanisms represented in its domain background context.

A.2.1 Temporal Coverage and Window Statistics

A step corresponds to an available numerical observation rather than a fixed calendar duration. We preserve the original observation calendar without interpolating missing dates; consequently,

[4] https://europeelects.eu/
[5] https://www.theguardian.com/
[6] https://www.bbc.com/
[7] https://www.telegraph.co.uk/
[8] https://magyarnemzet.hu/
[9] https://nepszava.hu/
[10] https://24.hu/
[11] https://www.spiegel.de/
[12] https://www.zeit.de/index
[13] https://www.faz.net/aktuell/
[14] https://www.ekathimerini.com/
[15] https://www.tovima.com/
[16] https://www.tanea.gr/

Table 5: Temporal coverage and window statistics by domain. History and horizon spans are measured in calendar days, while all windows contain 24 and 20 available observations, respectively.

Domain	Observation period	Candidate period	Anchor period	# Forecasting entities	Calendar span (days) History / Horizon
Energy	Jan 2, 2024 – Sep 9, 2026	Feb 5, 2024 – Aug 11, 2026	Apr 19, 2024 – Aug 10, 2026	1	31–35 / 28–33
Economy	Jan 2, 2024 – Sep 9, 2026	Feb 5, 2024 – Aug 11, 2026	May 29, 2024 – Aug 11, 2026	1	31–36 / 28–32
Politics	Jan 2, 2024 – Aug 21, 2026	Feb 7, 2024 – Jul 12, 2026	Mar 2, 2024 – Jul 5, 2026	10	32–42 / 30–40

Table 6: WIT benchmark composition by domain. Scenario counts are reported per anchor and in total; forecast-condition counts additionally include the scenario-free neutral condition.

Domain	Anchors	Factual / anchor	Alternative / anchor	Irrelevant / anchor	Scenario conditions	All forecast conditions
Energy	38	1	30	3	1,292	1,330
Economy	38	1	24	3	1,064	1,102
Politics	38	1	27	3	1,178	1,216
Total	**114**	**114 factual**	**3,078 alternatives**	**342 irrelevant**	**3,534**	**3,648**

the same 24-step history and 20-step forecast horizon can span different calendar durations across domains and anchors. Table 5 summarizes the observation, candidate, and final anchor periods, the number of forecasting entities, and the calendar spans of the history and forecast windows for each domain. The observation period covers the full range of valid observations; the candidate period covers forecast origins where complete 24-step histories and 20-step futures can be formed before screening; and the anchor period covers the forecast origins of the 38 anchors retained in the final benchmark.

A.2.2 Anchor and Scenario Composition

Table 6 summarizes the anchor-level composition of WIT. Each anchor contains one factual scenario, a domain-specific set of controlled alternative scenarios, and three irrelevant scenarios. Energy uses 30 controlled alternatives per anchor, Economy 24, and Politics 27. Across all anchors, this yields 3,078 controlled alternative anchor–scenario pairs, 342 irrelevant scenarios, and 114 factual scenarios.

Including the factual scenario, controlled alternatives, three irrelevant scenarios, and one neutral condition, each Energy, Economy, and Politics anchor requires 35, 29, and 32 forecasts, respectively. This gives 3,648 anchor-level forecasts per scenario-conditioned model and run; additional neutral forecasts for empirical matched cases are excluded from this count.

A.2.3 Controlled Scenario and Relation Inventory

The controlled scenario sets generate substantially more relational tests than the number of anchors alone would suggest. Table 7 reports the number of relations in each category. Each controlled alternative contributes a *Direction* relation; *Contrast* compares predefined opposing scenario pairs; *Intensity* compares adjacent levels within ordered scenario-strength sequences; and each irrelevant scenario contributes a *Restraint* relation.

Across the three domains, WIT contains 3,078 Direction, 1,482 Contrast, 2,052 Intensity, and 342 Restraint relations, for a total of 6,954 relational tests. Because the domain-level scenario sets differ in size, each Energy, Economy, and Politics anchor contributes 68, 55, and 60 relations, respectively.

A.2.4 Empirical Matched-Set Coverage

Grounded Comparison uses a subset of anchor–scenario pairs with at least three empirical matched cases. The current release contains 20 supported pairs: 3 in Energy, 11 in Economy, and 6 in Politics, with 62 matched-case assignments in total and three to four cases per supported pair (Table 8).

Table 7: Relational Comparison inventory by domain. Counts aggregate predefined relations over all 38 anchors; values in parentheses show relations per anchor.

Domain	Direction	Contrast	Intensity	Restraint	Total relations	Relations / anchor
Energy	1,140 (30)	570 (15)	760 (20)	114 (3)	2,584	68
Economy	912 (24)	456 (12)	608 (16)	114 (3)	2,090	55
Politics	1,026 (27)	456 (12)	684 (18)	114 (3)	2,280	60
Total	**3,078**	**1,482**	**2,052**	**342**	**6,954**	–

Table 8: Empirical matched-set support for Grounded Comparison. Supported pairs have at least three empirical matched cases. Total matched cases sum pool sizes, counting reuse across pairs.

Domain	Supported pairs	Total matched cases	Matched cases / pair (min / median / max)
Energy	3	10	3 / 3 / 4
Economy	11	33	3 / 3 / 3
Politics	6	19	3 / 3 / 4
Total	**20**	**62**	–

This population is smaller than the controlled relational inventory because Grounded Comparison is restricted to pairs with sufficient empirical support. Grounded Curve Error is therefore reported separately by domain, while Grounded Scenario Gain is macro-averaged with equal domain weight so that Economy's larger number of supported pairs does not dominate the aggregate result.

B Benchmark Construction Protocol

This section details the operational steps of the reusable construction framework in Section 3. The same procedure is applied across domains, with domain-specific numerical and textual sources used to instantiate each benchmark component.

B.1 Candidate Windows and Forecasting States

Candidate windows are generated with stride one over consecutive available observations on each target's original observation calendar, without interpolation. Each candidate i contains a 24-step numerical history $\mathbf{x}_i$ and a 20-step observed future $\mathbf{y}_i$, and is retained only when both windows can be fully formed from available observations. The corresponding forecasting state is $z_i = (\mathbf{x}_i, \mathcal{C}_D)$, where $\mathcal{C}_D$ is the fixed domain background context. Thus, $L = 24$ and $H = 20$ refer to numbers of available observations rather than fixed calendar durations.

B.2 Domain Background Context Construction

For each domain D, the domain background context $\mathcal{C}_D$ is constructed from reference material using the same procedure, with all sources published no later than December 31, 2023. It defines the forecasting target and a compact mechanism map describing general external drivers, their supported qualitative effects, applicability conditions, and limitations. Information tied to specific benchmark-period events, including timestamped outlooks and anchor-specific events or conditions, is excluded.

Candidate histories and futures are drawn only from observations from January 2024 onward. These temporal restrictions serve distinct purposes: the background cutoff separates general domain knowledge from benchmark-period events, while the observation cutoff reduces overlap with the reported knowledge cutoffs of the evaluated language models.

B.3 Reference-Guided Candidate Screening

We use Chronos-2 (Ansari et al., 2025), TimesFM-3 (Jain &Sen, 2026), and TiRex-2 (Podest et al., 2026) as frozen history-only construction references. Screening identifies candidates whose observed futures show consistent and substantial deviations from the calibrated numerical expectations of all three references, without using target-unit-specific thresholds.

B.3.1 Construction-Only Calibration Pool

For each domain D, let $\mathcal{K}_D$ denote a construction-only calibration pool fixed before candidate screening and kept separate from released anchors and empirical matched cases. Because stride-one windows can share much of the same history and future, calibration is performed separately for each candidate using only non-overlapping calibration windows.

For candidate i, let I_i denote the calendar interval spanning its history and future. Its admissible calibration pool is

$$\mathcal{K}_D^{(-i)} = \{c \in \mathcal{K}_D : I_c \cap I_i = \varnothing\}. \tag{12}$$

Thus, no calibration window whose history–future interval overlaps that of candidate i contributes to its calibration statistics.

B.3.2 Cross-Fitted Reference-Relative Deviations

For reference model m and forecast lead h, define the calibration residual

$$e_{c,h}^{(m)} = y_{c,h} - b_{c,h}^{(m)}, \tag{13}$$

where $b_{c,h}^{(m)}$ is the frozen history-only forecast for calibration window c. Using only $c \in \mathcal{K}_D^{(-i)}$, we estimate a robust center separately for each forecast step h

$$\mu_{D,m,h}^{(-i)} = \mathrm{median}_{c \in \mathcal{K}_D^{(-i)}}\, e_{c,h}^{(m)} \tag{14}$$

and the corresponding raw robust scale

$$\widetilde{\sigma}_{D,m,h}^{(-i)} = 1.4826\ \mathrm{MAD}_{c \in \mathcal{K}_D^{(-i)}}\, e_{c,h}^{(m)}, \tag{15}$$

where MAD denotes the median absolute deviation from the median, providing a robust estimate of residual dispersion that is less sensitive to extreme forecast errors than the standard deviation; the factor 1.4826 rescales it to the standard-deviation scale under a Gaussian distribution.

Zero-scale safeguard. The MAD at a given forecast step can be zero when the admissible residuals show insufficient variation. Rather than introducing an arbitrary numerical constant, we first borrow scale information from other forecast leads of the same reference model and domain. We define

$$\sigma_{D,m,\mathrm{pool}}^{(-i)} = 1.4826\ \mathrm{MAD}\left(\left\{e_{c,h}^{(m)} - \mu_{D,m,h}^{(-i)} : c \in \mathcal{K}_D^{(-i)},\ h = 1, \ldots, H\right\}\right). \tag{16}$$

The effective scale is then

$$\sigma_{D,m,h}^{(-i)} = \begin{cases} \widetilde{\sigma}_{D,m,h}^{(-i)}, & \widetilde{\sigma}_{D,m,h}^{(-i)} > 0, \\ \sigma_{D,m,\mathrm{pool}}^{(-i)}, & \widetilde{\sigma}_{D,m,h}^{(-i)} = 0 \ \text{ and } \ \sigma_{D,m,\mathrm{pool}}^{(-i)} > 0, \\ \delta_D, & \text{otherwise}, \end{cases} \tag{17}$$

where δ_D is the documented numerical resolution of the target. This fallback preserves the center estimated for each forecast step while preventing division by zero or an arbitrarily small scale.

The candidate residual is standardized as

$$z_{i,h}^{(m)} = \frac{y_{i,h} - b_{i,h}^{(m)} - \mu_{D,m,h}^{(-i)}}{\sigma_{D,m,h}^{(-i)}}. \tag{18}$$

This expresses each deviation relative to the typical forecasting error of the corresponding reference model at the same forecast step, rather than in the native units of the target.

We determine the dominant direction of the candidate deviation by

$$p_i = \operatorname{sign}\left(\operatorname{median}_{m,h} z_{i,h}^{(m)}\right), \tag{19}$$

and define the lead-wise cross-model consensus as

$$\gamma_{i,h} = \min_m p_i z_{i,h}^{(m)}. \tag{20}$$

Because the minimum is taken across reference models, $\gamma_{i,h} > 0$ only when all references deviate from the observed future in the same direction at lead h.

We summarize the persistence and magnitude of these consensus deviations using

$$S_i = Q_{0.25}\left(\gamma_{i,1}, \ldots, \gamma_{i,H}\right). \tag{21}$$

The lower quartile prevents a small number of extreme forecast errors from determining candidate eligibility. A positive S_i requires same-direction cross-model consensus over most of the forecast horizon, while larger values indicate departures that are large relative to the references' usual forecasting errors.

Cross-fitted percentile calibration. Calibration windows are scored using the same leave-overlap-out procedure. For each $c \in \mathcal{K}_D$, its score S_c is computed using calibration statistics that exclude all windows whose intervals overlap I_c. For candidate i, only scores from calibration windows that do not overlap I_i are used to form its reference distribution:

$$\widehat{F}_i(s) = \frac{1}{|\mathcal{K}_D^{(-i)}|} \sum_{c \in \mathcal{K}_D^{(-i)}} \mathbf{1}[S_c \le s]. \tag{22}$$

The calibrated departure percentile is then

$$\pi_i = \widehat{F}_i(S_i). \tag{23}$$

Thus, π_i measures how unusually large the candidate's persistent cross-model deviation is relative to admissible construction-only calibration windows.

B.3.3 Candidate Eligibility

A candidate is numerically eligible only if $S_i > 0$ and $\pi_i > \kappa$. The current release uses the common threshold $\kappa = 0.90$, so an eligible candidate must exceed the 90th percentile of its admissible cross-fitted calibration distribution in addition to exhibiting a positive persistent consensus departure. The value of κ controls screening stringency rather than statistical significance and is fixed before anchor selection and applied identically across domains.

Together, the two conditions require both persistent same-direction agreement across the construction references and an unusually large deviation relative to the domain's calibrated reference-error distribution. Because the procedure operates first in standardized residual space and then in percentile space, the same screening rule can be applied to forecasting targets with different units and numerical scales.

B.4 Factual Evidence-to-Scenario Construction

B.4.1 Source-Supported Factual Evidence

For each numerically eligible candidate, we retrieve time-stamped source evidence relevant to the forecast interval and to mechanisms in the fixed domain background context $\mathcal{C}_D$. Each evidence record retains the source, timing, affected entity or scope, and supporting passages. The source establishes what condition occurred and under what circumstances, while $\mathcal{C}_D$ determines whether the condition maps to a supported forecasting mechanism.

B.4.2 OUTCOME-BLIND EVENT EXTRACTION

Conditions are extracted without access to the realized target trajectory, reference-relative deviations, calibrated departure score, or preferred response direction. Only source-supported conditions are retained, together with their timing, scope, mechanism, and applicability requirements. Unsupported, conflicting, or ambiguous claims remain unresolved rather than being inferred from the numerical trajectory.

B.4.3 OUTCOME-BLIND FACTUAL SCENARIO COMPOSITION

Verified conditions are selected using only source support, target relevance, temporal relevance, and mechanism applicability. The selected conditions are condensed into a compact factual scenario that preserves information required to interpret the mechanism, including relevant timing, scope, roles, and uncertainty, while removing unnecessary source-specific or identifying details.

The model-facing scenario contains no realized target information, reference forecasts, calibrated residuals, quantified target effects, or explicit forecasting instructions. Any qualitative alignment between the fixed scenario and the realized reference-relative deviation is assessed only afterward and does not affect scenario wording or sample inclusion.

B.4.4 SOURCE AND SEMANTIC REVIEW

A separate review checks the fixed scenario against the original source evidence and $\mathcal{C}_D$. Each substantive claim must be source-supported, consistent with the reported timing, scope, status, and uncertainty, and satisfy the applicability requirements of its assigned mechanism. Unsupported claims are not repaired using outside knowledge or realized target information; scenarios requiring substantive correction are revised and reviewed again.

B.4.5 SAMPLE ACCEPTANCE

A candidate proceeds to anchor formation only after passing both numerical screening and factual-evidence review. Acceptance requires at least one supported, forecast-relevant condition from which a valid factual scenario can be composed; agreement with the realized reference-relative deviation is not required. Candidates with unresolved occurrence, timing, scope, or mechanism applicability do not proceed to anchor formation.

B.5 FINAL ANCHOR SELECTION

Only candidates that pass both numerical screening and factual-evidence review are eligible for final anchor selection. Once eligibility is established, candidates are not further ranked by realized future magnitude, reference-relative deviation strength, S_i, or π_i. Final selection instead aims to provide broad temporal coverage while limiting near-duplicate forecasting states.

Temporal coverage and redundancy control. Redundancy is controlled within each forecasting entity using the distance between forecast origins in the original sequence of available observations rather than calendar time. We define the minimum separation as $\Delta = \lceil \rho_{\text{sep}} L \rceil$ with $\rho_{\text{sep}} = 0.5$, where L is the history length. For the current release, $L = 24$, giving $\Delta = 12$ available observation steps across all domains. This ensures that selected anchors from the same forecasting entity share at most half of their numerical history, reducing highly redundant forecasting states while retaining broad temporal coverage. Because separation is measured in available observations, the same rule applies to regularly and irregularly observed target series without introducing domain-specific calendar-time thresholds.

To make selection deterministic, the evaluation period is divided into equally spaced temporal strata, and eligible candidates closest to the center of each stratum are selected subject to the separation rule, with ties resolved by the earlier forecast origin.

For Politics, the separation rule is applied independently within each party-specific series; candidates from different parties are not treated as redundant solely because their forecast origins are temporally close.

Final anchor selection is independent of empirical matched-set availability. The Factual and Relational anchor population is therefore determined without considering whether an anchor–scenario pair is subsequently supported for Grounded Comparison.

B.6 Controlled Scenario Set Construction

Controlled scenarios are constructed once for each domain from the fixed domain background context $\mathcal{C}_D$ and reused across all anchors. Each scenario is self-contained: the information needed to understand how the scenario relates to the forecasting target, including relevant roles, relationships, timing, scope, and applicability conditions, is stated explicitly rather than inferred from a particular anchor. Only relations that can be directly determined from the scenario text and $\mathcal{C}_D$ are retained.

For mechanisms supported by $\mathcal{C}_D$, we construct hypothetical conditions for three relation types. ***Direction*** specifies the expected sign of the scenario-induced revision relative to the neutral condition. ***Contrast*** pairs conditions with opposing implications under the same mechanism while keeping timing, scope, and strength comparable. ***Intensity*** compares stronger and weaker versions of the same condition by varying a single strength attribute while holding other relevant attributes fixed. ***Restraint*** is constructed separately using scenarios unrelated to the forecasting target under $\mathcal{C}_D$, for which no meaningful revision is expected.

Relations requiring unstated or anchor-specific assumptions are not included. Scenario texts, relation labels, and the attributes held fixed or varied within each relation are reviewed and fixed before model evaluation.

B.7 Empirical Matched-Set Construction

For each anchor–scenario pair (i, s), empirical matching considers cases with (i) a similar numerical history, (ii) similar history-only forecasts from the frozen construction references, and (iii) source evidence that the same scenario condition actually occurred in comparable forecasting settings. Numerical and semantic matching are performed independently, and realized futures are not used during matching.

Numerical matching. To reduce dependence on absolute target levels, histories and construction-reference forecasts are expressed relative to each case's last observed value and scaled by a construction-only robust domain scale ω_D. For a single-series domain, ω_D is the median nonzero change between consecutive available observations; for a multi-entity domain, this scale is first computed separately for each entity and then aggregated by the median across entities. For anchor i, we define

$$\widetilde{x}_{i,\ell} = \frac{x_{i,\ell} - x_{i,L}}{\omega_D}, \qquad \widetilde{b}^{(m)}_{i,h} = \frac{b^{(m)}_{i,h} - x_{i,L}}{\omega_D}, \tag{24}$$

where $b^{(m)}_{i,h}$ denotes the history-only forecast from construction reference m.

The history distance between anchor i and candidate case j is

$$d_x(i,j) = \sqrt{\frac{1}{L}\sum_{\ell=1}^{L} \left(\widetilde{x}_{i,\ell} - \widetilde{x}_{j,\ell}\right)^2}, \tag{25}$$

and the corresponding reference-forecast distance is computed separately for each construction reference:

$$d^{(m)}_b(i,j) = \sqrt{\frac{1}{H}\sum_{h=1}^{H} \left(\widetilde{b}^{(m)}_{i,h} - \widetilde{b}^{(m)}_{j,h}\right)^2}. \tag{26}$$

To avoid domain-specific raw-distance thresholds, these distances are converted to percentile ranks. For each anchor i, candidate j is ranked relative to the corresponding distances between i and the non-overlapping same-domain construction windows in $\mathcal{K}_D^{(-i)}$. Let $r_x(i,j)$ and $r^{(m)}_b(i,j)$ denote the resulting history and reference-forecast distance percentiles, with lower values indicating greater

similarity. The current release uses a common threshold $\alpha = 0.10$, and a candidate passes numerical matching only if

$$r_x(i,j) \le \alpha, \qquad \max_m r_b^{(m)}(i,j) \le \alpha. \tag{27}$$

Thus, a matched case must lie among the closest 10% of admissible cases in numerical history and satisfy the same similarity criterion for every construction reference.

Semantic matching and final selection. Numerically admissible cases are retained only when source evidence independently verifies that the same scenario condition actually occurred and was relevant to the target under comparable timing, scope, and roles. Self-matches and repeated instances of the same underlying event are removed, and temporal redundancy within each forecasting entity is controlled using the observation-step separation rule in Appendix B.5.

Among cases satisfying all numerical and semantic criteria, we define

$$r_{\mathrm{num}}(i,j) = \frac{1}{2} r_x(i,j) + \frac{1}{2M} \sum_{m=1}^{M} r_b^{(m)}(i,j), \tag{28}$$

where M is the number of construction references. Admissible cases are ordered by $r_{\mathrm{num}}(i,j)$, with ties resolved by earlier forecast origin. Up to four cases are retained for each anchor–scenario pair, and pairs with fewer than three admissible cases are marked unsupported.

Matched cases are not required to satisfy the persistent-deviation criterion used for anchor screening, preventing the empirical reference set from being preselected for unusually large realized responses.

B.8 Leakage Controls, Freezing, and Audit Trail

WIT completes benchmark construction before model evaluation, with all evaluation components fixed in advance. The domain background context uses only reference material published by December 31, 2023, while candidate windows are drawn from post-2023 observations. Factual evidence extraction and scenario composition are outcome-blind, and realized matched-case futures are not used when constructing empirical matched sets. No outputs from subsequently evaluated models are used during benchmark construction.

All benchmark components that determine evaluation outcomes—the final anchor set, factual and controlled scenario texts, relation labels, supported anchor–scenario pairs, empirical matched sets, and scoring rules—are fixed before model evaluation. The construction archive retains the source, evidence, and decision records required to trace numerical screening, annotation, and matching, while keeping this information separate from model-facing inputs.

Each benchmark version is immutable once released. Any later change resulting from source, evidence, or construction review is issued as a new version rather than modifying the existing evaluation specification.

C Construction Prompts and Model Inputs

C.1 Construction Prompt Templates

C.1.1 Domain Background Context Construction

The same prompt structure is used across domains to construct the fixed domain background context $\mathcal{C}_D$ from pre-cutoff reference material. The prompt separates general, reusable domain mechanisms from benchmark-period events and requires each mechanism to retain its applicability conditions and uncertainty (Table 9).

C.1.2 Event Extraction

The *event extraction* prompt identifies source-supported conditions that are relevant to the forecast interval and correspond to mechanisms in $\mathcal{C}_D$. It records their timing, scope, status, mechanism, and applicability while preventing inference from the target trajectory, model forecasts, or other

construction-derived signals. Unsupported or unresolved conditions are explicitly retained rather than inferred (Table 10).

C.1.3 FACTUAL SCENARIO COMPOSITION

The *scenario composition* prompt selects up to two verified conditions based only on source support, target and temporal relevance, and mechanism applicability, and compresses them into a concise, mechanism-preserving factual scenario. It preserves necessary qualifications while excluding realized target information, other construction-derived signals, unsupported assumptions, and explicit forecasting guidance (Table 11).

C.1.4 SOURCE AND SEMANTIC REVIEW

The *source and semantic review* prompt audits the proposed factual scenario against the original evidence and $\mathcal{C}_D$. It verifies source support, timing and scope, mechanism applicability, preservation of uncertainty, and the absence of leaked outcome information or unsupported causal interpretation (Table 12). Scenarios requiring substantive correction must be revised and reviewed again before acceptance.

C.2 MODEL-FACING INPUT EXAMPLES

C.2.1 DOMAIN BACKGROUND CONTEXT

The domain background context provides fixed, domain-level knowledge shared across all anchors within each domain. It defines the forecasting target and summarizes general mechanisms through which external conditions may affect it. For Energy, these mechanisms cover weather-driven demand, production, pipelines, storage, LNG and trade, and industry or fuel substitution. For Economy, they cover monetary policy and relative interest rates, relative economic conditions, global risk sentiment, and foreign developments and capital flows. For Politics, they cover economic and public-service conditions, government performance, leadership and party unity, policy positions and issue salience, and political competition. Each background also states relevant applicability conditions and qualifications rather than assuming that the same directional effect holds in every setting.

The background contains no anchor-specific events, future scenarios, or other time-varying information from the benchmark period. The full domain background contexts $\mathcal{C}_D$ for all three domains are provided in Tables 13–15.

C.2.2 FACTUAL SCENARIO

The factual scenario summarizes source-supported external conditions that occurred or remained relevant during the forecast horizon. Across Energy, Economy, and Politics, it preserves the timing, scope, roles, and other information needed to interpret the corresponding mechanism in $\mathcal{C}_D$, while excluding realized target outcomes and explicit forecasting guidance. The scenario is constructed from the verified textual evidence and provided as the factual condition for forecasting. Examples from each domain are provided in Table 16.

C.2.3 CONTROLLED ALTERNATIVE SCENARIOS

Controlled alternative scenarios describe plausible future conditions under the same forecasting state, varying only the scenario condition. They are constructed from mechanisms in the corresponding domain background context and reused across anchors to support controlled comparisons of scenario responses. Domain-specific examples are provided in Tables 17–19.

C.2.4 IRRELEVANT SCENARIO

Irrelevant scenarios describe conditions unrelated to the forecasting target under the corresponding domain background context. They test whether the model avoids unnecessary forecast revisions when the provided scenario is not relevant to the target. An example is provided in Table 20.

C.3 Evaluation Prompt Templates

We provide the same numerical and textual information across models while adapting it to each model's input interface. For directly prompted LLMs, the numerical history and textual inputs are combined in a single prompt. For context-aware forecasting systems, the same textual content is provided through their designated context inputs while preserving the original model interface. The following templates specify these input formats.

C.3.1 Directly Prompted LLMs

Directly prompted LLMs receive the numerical history, domain background context, and scenario in a single prompt. They are instructed to use scenario information when relevant and to rely on the numerical history and domain background when the scenario is unrelated to the target. The output is restricted to a JSON array containing exactly the required number of forecast values, with no additional explanation. The prompt template is provided in Table 21.

C.3.2 Context-aware Forecasting Systems

For context-aware forecasting systems, the numerical history is provided through each model's native time-series input, while the domain background context and scenario are passed through the designated textual input using a minimal template. For DoubleCast (Zheng et al., 2026), this follows the simple context format used in CAF-7M, while MoiraiAgent (Yan et al., 2026) processes the same textual content through its GPT-4o backbone. We do not add an explicit instruction to ignore irrelevant scenarios, so that relevance handling remains part of each system's native context-processing behavior. The textual input template is provided in Table 22.

Table 9: Domain background context construction prompt.

```
ROLE
Construct a fixed domain background for a scenario-conditioned forecasting benchmark.

INPUT
- Target definition and units
- Background cutoff date
- Pre-cutoff reference passages

TASK
Write:
1. a concise definition of the forecasting target, and
2. a compact set of general driver mechanisms that can affect it.

For each mechanism, describe:
- the relevant condition or driver,
- its conditional upward or downward pressure where supported,
- assumptions, applicability conditions, and limitations required for that
interpretation.

RULES
- Use only the supplied pre-cutoff reference material.
- Include only general mechanisms that can be reused across benchmark episodes.
- Exclude benchmark episodes, anchor-specific entities or dates, post-cutoff events,
  dated forecasts, contemporary settings, and episode-specific expectations.
- Do not predict the target or assign numerical effect sizes.
- Do not infer that any mechanism actually occurs in a benchmark episode.
- Preserve uncertainty and conditionality; do not turn context-dependent relations
  into universal directional rules.
- Omit mechanisms not supported by the supplied sources.

OUTPUT
background_text,
mechanism_map: mechanism, supported directions, required assumptions,
               limitations, supporting source IDs, unresolved_items
```

Table 10: Event extraction prompt.

```
ROLE
Extract source-supported conditions relevant to scenario annotation.

INPUT
- Target and fixed domain background
- Forecast interval
- Relevant entity, geography or scope, and role or relation if applicable
- Dated source passages

TASK
Identify distinct events or conditions that:
1. are explicitly supported by the supplied sources,
2. overlap or remain relevant during the forecast interval, and
3. correspond to a mechanism in the fixed domain background.

For each condition, record:
- factual condition,
- timing and scope,
- event status (observed, expected, announced, alleged, resolved, etc.),
- applicable mechanism and required assumptions,
- conflicting evidence or unresolved information.

RULES
- Use only the supplied sources.
- Do not use outside knowledge.
- Do not infer events from the target trajectory, model forecasts, reference-relative
  direction, or calibrated departure score.
- Do not infer that a condition caused, dominated, or explains the target movement.
- Preserve uncertainty and distinguish occurrence from expectation, announcement,
  allegation, and resolution.
- Mark missing applicability conditions explicitly rather than filling them in.
- If no usable condition is supported, return INSUFFICIENT.

OUTPUT
evidence_status, extracted_events, unresolved_items
```

Table 11: Scenario composition prompt.

```
ROLE
Convert source-supported event records into a concise factual scenario.

INPUT
- Target and fixed domain background
- Source-supported event records
- Forecast interval

TASK
Select at most two conditions using only evidence-based criteria:
1. strength of source support,
2. relevance to the forecasting target,
3. temporal relevance to the forecast interval, and
4. applicability of the corresponding domain mechanism.

Express the selected conditions as one or two concise, mechanism-preserving sentences.

RULES
- Use only the supplied event records.
- Do not use the realized target trajectory, construction forecasts,
  reference-relative direction, or calibrated departure score.
- Preserve qualifications required for timing, scope, status, uncertainty,
  and mechanism validity.
- Remove identifying details that are unnecessary for interpretation while
  preserving roles, relationships, scope, and other attributes required for
  the mechanism to remain meaningful.
- Do not reveal target values, realized target movement, numerical effect sizes,
  or instructions about how the forecast should change.
- Do not introduce unsupported assumptions or claim that a selected condition
  caused, dominated, or fully explained the realized future.
- If no sufficiently supported and forecast-relevant condition is available,
  return HOLD_SAMPLE.

OUTPUT
status, scenario_text, selected_event_ids, mechanism_ids,
claim_map, retained_qualifications, hold_reason
```

Table 12: Source and semantic review prompt.

```
ROLE
Audit a proposed factual scenario against its original evidence and the fixed
domain specification.

INPUT
- Fixed domain background and mechanism map
- Proposed scenario and claim-to-source mapping
- Source-supported event records and original source passages
- Forecast interval
- Relevant entity, geography or scope, and role or relation if applicable

CHECK
For every substantive claim, verify:
1. direct support from the cited source,
2. correct timing, scope, entity, and event status,
3. applicability of the assigned mechanism and its required assumptions,
4. preservation of uncertainty and necessary qualifications,
5. fidelity after compression or de-identification,
6. absence of realized target information, forecasting instructions,
   or unsupported causal interpretation.

RULES
- Use only the supplied evidence and fixed domain specification.
- Do not use outside knowledge.
- Do not use or infer validity from the realized target trajectory,
  construction forecasts, reference-relative direction, or calibrated departure score.
- Do not repair unsupported claims with plausible assumptions.
- Any unsupported claim necessary for the scenario's interpretation requires
HOLD_SAMPLE.
- A revised scenario must be reviewed again before acceptance.

OUTPUT
verdict: PASS or HOLD_SAMPLE
claim_verdicts, required_corrections, unresolved_items
```

Table 13: Domain background context $\mathcal{C}$ in the Energy domain.

```
The target is the Henry Hub natural gas spot price (USD/MMBtu). Higher demand or
less available gas generally pushes prices up; lower demand or more available gas
generally pushes prices down.

  - Weather demand: Cold winter weather increases gas use for heating. Hot summer
  weather increases electricity use for air conditioning, which can increase gas
  use at power plants. Warmer winters or cooler summers reduce these demands.

  - Production: Freezing weather and equipment shutdowns can reduce gas production.
  Production restarts or growth add supply. Storms can also reduce gas use by
  disrupting businesses and electricity demand, partly offsetting production
  losses.

  - Pipelines: A blocked pipeline can leave consumers short of gas while trapping
  excess gas near producers. Prices may rise where gas cannot arrive and fall where
  gas cannot leave. The effect on Henry Hub depends on how the affected pipeline
  connects to the hub.

  - Storage: Less stored gas than usual for the season leaves a smaller reserve to
  meet demand; more stored gas provides a larger reserve. Withdrawals above
  expectations or injections below expectations suggest a tighter supply-demand
  balance; the reverse suggests a looser balance.

  - LNG and trade: More LNG exports leave less U.S. gas available for domestic use.
  Export-terminal shutdowns reduce exports and leave more gas at home; restarts
  reverse this effect. Overseas events matter when they affect U.S. gas imports or
  exports, within the limits of available infrastructure.

  - Industry and fuel substitution: Higher industrial activity increases gas use.
  When nuclear or hydropower output falls, gas-fired plants may burn more gas to
  replace it. More renewable generation or switching to other fuels can reduce gas
  use.
```

Table 14: Domain background context $\mathcal{C}$ in the Economy domain.

```
The target is the Nominal Broad U.S. Dollar Index (DTWEXBGS), a trade-weighted
measure of the U.S. dollar against major U.S. trading partners. A higher index
means a stronger dollar; a lower index means a weaker dollar.

  - Monetary policy and relative interest rates: Higher expected U.S. interest
  rates relative to other economies generally support the dollar, while lower
  relative rates or faster expected U.S. easing tend to weaken it.

  - Relative economic conditions: Stronger U.S. growth or economic data relative to
  major trading partners generally support the dollar, while weaker relative
  conditions tend to weigh on it. Inflation matters mainly through its effects on
  expected monetary policy and real returns.

  - Global risk sentiment: Financial stress, geopolitical uncertainty, or weaker
  risk appetite often increase demand for liquid U.S. dollar assets and strengthen
  the dollar. Improving global risk sentiment can reduce this safe-haven demand.

  - Foreign developments and capital flows: Stronger demand for U.S. assets or
  weaker monetary and economic conditions abroad generally support the dollar,
  while capital shifting toward foreign assets can weaken it. Effects are larger
  for currencies with greater weights in the broad dollar index.
```

Table 15: Domain background context $\mathcal{C}$ in the Politics domain.

```
The target is the selected party's reported polling support in its country,
measured as a percentage. A higher value means greater support for that party; a
lower value means less support. Effects depend on the party's role and on which
parties voters credit, blame, or view as credible alternatives.

  - Economic conditions and public services: Improving living standards,
  employment, housing, or public services can benefit parties credited with those
  outcomes, while deteriorating conditions can hurt parties held responsible.

  - Government performance and policy delivery: Effective administration and
  delivery of valued policies can strengthen support, while policy failures,
  unpopular decisions, or broken commitments can weaken it. Responsibility may be
  shared unevenly in coalition governments.

  - Leadership, integrity, and party unity: Favorable evaluations of party
  leadership, competence, and cohesion can strengthen support, while leadership
  disputes, defections, scandals, or perceived integrity failures can weaken it.

  - Policy positions and issue salience: Parties can gain when issues on which they
  are viewed as credible become more important and their positions align with voter
  priorities. They can lose when salient positions conflict with those priorities
  or are viewed as ineffective.

  - Political competition and alternatives: Stronger campaigns, alliances, or
  coordination can consolidate support, while credible rivals, fragmentation, or
  shifts toward competing parties can reduce it. Dissatisfaction with one party
  does not necessarily benefit any particular alternative.
```

Table 16: Example factual scenario s^* for each domain.

```
(Energy) -> Weather-driven demand increase and production disruption mechanism
Unusually cold weather across the U.S. increased heating demand and withdrawals
from gas storage, while freeze-offs reduced natural gas production.

(Economy) -> Monetary policy and relative interest rates mechanism
The Federal Reserve kept interest rates elevated while the European Central Bank
began lowering its policy rates.

(Politics) -> leadership disputes and party cohesion mechanism
Public disagreements between the party leadership and its own members exposed
internal divisions.
```

Table 17: Examples of alternative scenarios in the Energy domain.

```
(Weather Demand)
- Unusually mild weather reduces heating demand across major U.S. population
centers.

(Production)
- Equipment failures at gas-processing plants reduce available supply until repairs
are completed later.

(Pipelines)
- A pipeline outage restricts gas deliveries to the Henry Hub area, reducing local
availability until repairs are completed.

(Storage)
- A weekly storage report shows larger-than-expected withdrawals, bringing
inventories below their seasonal average.

(LNG and trade)
- An LNG export terminal resumes operations and gradually increases the gas it
draws from U.S. pipelines.

(Industry and fuel substitution)
- Stronger wind generation replaces some gas-fired electricity generation, reducing
gas use at power plants.
```

Table 18: Examples of alternative scenarios in the Economy domain.

```
(Monetary policy and relative interest rates)
- Expected U.S. interest rates rise relative to those in major trading partners as
the Federal Reserve signals a slower pace of easing.

(Relative economic conditions)
- U.S. hiring and consumer spending weaken while economic activity in major trading
partners remains resilient.

(Global risk sentiment)
- Heightened financial stress increases international investors' demand for liquid
U.S. dollar assets.

(Foreign developments and capital flows)
- Investors shift funds from U.S. assets toward foreign assets, selling dollars to
purchase foreign currencies.
```

Table 19: Examples of alternative scenarios in the Politics domain.

```
(Economic conditions and public services)
- Household living costs rise and public services deteriorate, with voters holding
the party responsible.

(Government performance and policy delivery)
- A public-service reform championed by the party is successfully implemented, and
voters credit the party for the improvement.

(Leadership, integrity, and party unity)
- Disputes among senior party figures intensify, exposing public disagreements over
leadership and policy.

(Policy positions and issue salience)
- Public attention shifts toward an issue on which voters regard the party as
competent and favor its policy proposals.

(Political competition and alternatives)
- A direct rival forms an electoral alliance around a shared platform that appeals
to the party's existing voter base.
```

Table 20: Example of an irrelevant scenario.

```
A university library updates its online manuscript catalog, correcting publication
dates and adding subject tags to help researchers find historical documents more
easily.
```

Table 21: Prompt template for directly prompted LLMs.

```
Forecast the target time series using its history, domain background, and relevant
scenario information. If the scenario is unrelated to the target, rely on the
historical time series and domain background.

Domain background: {domain_background}

Historical time series: {history_timeseries}

Scenario: {scenario_text}

Forecast the next {prediction_horizon} observations.
Output only a JSON array of exactly {prediction_horizon} numbers, ordered from
earliest to latest, in the same units and scale as the history. Do not include
explanations or Markdown.
```

Table 22: Prompt template for context-aware forecasting systems.

```
Domain background: {domain_background}
Scenario: {scenario_text}
```

C.3.3 WIT Evaluation Example

Figure 5 provides a concrete example of WIT's Factual, Relational, and Grounded comparisons using the Economy domain.

D Additional Evaluation Details

D.1 Evaluation Populations and Aggregation

For Relational Comparison, relation outcomes are first averaged within each anchor, anchors are weighted equally within each relation category, and the four category scores are weighted equally in Overall SLA. This prevents anchors or categories with more predefined relations from having greater influence on the aggregate score.

For Grounded Comparison, supported anchor–scenario pairs are weighted equally within each domain. GCE is reported separately by domain, while GSG is macro-averaged with equal domain weight so that domains with more supported pairs do not dominate the overall score.

D.2 Factual Comparison Details

Within each domain, FSG is computed from the ratio of aggregate factual and neutral errors rather than by averaging anchor-specific percentage changes. Overall FSG is then macro-averaged with equal domain weight. CWR is the strict share of anchors satisfying $e_i^{s_i^*} < e_i^0$, with ties not counted as improvements, and is likewise macro-averaged across domains.

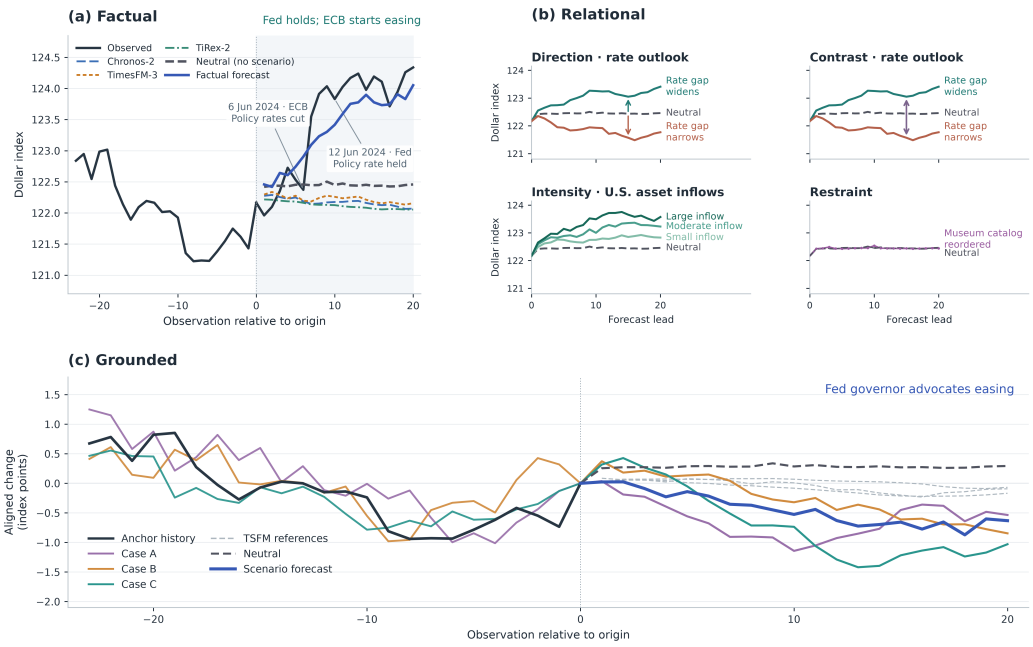


Figure 5: Example of WIT's three comparisons in the Economy domain. (a) **Factual**: the observed history and future, timestamped supporting evidence, and forecasts under the factual and neutral conditions, together with numerical TSFM references. (b) **Relational**: controlled alternatives illustrate Direction and Contrast through opposing relative-rate conditions, Intensity through progressively stronger U.S. asset inflows with net dollar purchases, and Restraint through an irrelevant scenario. (c) **Grounded**: comparable real cases provide empirical references for evaluating the neutral-relative revision induced by the alternative scenario.

D.3 Relational Comparison Details

D.3.1 Relation Eligibility and Scoring

Each predefined relation is evaluated only when all scenario conditions required by its construction are available for the anchor. Direction is evaluated for individual controlled scenarios, Contrast for predefined opposing pairs, Intensity for adjacent levels of an ordered scenario sequence, and Restraint for irrelevant scenarios. Each relation produces a binary outcome according to the criteria in Section 4 and the practical tolerances defined below.

D.3.2 Practical Tolerances

The relational criteria use practical tolerances to avoid treating numerically negligible forecast changes as meaningful responses. For Direction, Contrast, and Intensity, η_D specifies the minimum meaningful revision; for Restraint, τ_D specifies the largest revision treated as negligible. These are practical-effect thresholds rather than statistical significance thresholds.

Reusable scale-based rule. Because target units and numerical variation differ across domains, we derive the tolerances from a common scale-based rule. For forecasting entity e in domain D, let

$$\mathcal{U}_{D,e} = \left\{ |x_{e,t_k} - x_{e,t_{k-1}}| : k = 2, \ldots, T,\ x_{e,t_k} \neq x_{e,t_{k-1}} \right\}, \tag{29}$$

and define

$$v_{D,e} = \operatorname{median} \mathcal{U}_{D,e}. \tag{30}$$

Thus, $v_{D,e}$ represents a typical non-zero change between consecutive available observations. For a single-entity domain, $v_D = v_{D,e}$; for a multi-entity domain,

$$v_D = \operatorname{median}_e v_{D,e}. \tag{31}$$

Table 23: Sensitivity of Relational Comparison to the practical tolerance. Scores are macro-averaged across the three domains. $0.5\times$, $1\times$, and $2\times$ scale both η_D and τ_D relative to the default setting.

Tolerance	Model	Direction	Contrast	Intensity	Restraint	Overall
$0.5\times$	Mistral (7B)	73.45	70.95	57.85	66.96	67.30
	Llama-3 (70B)	82.35	80.05	68.65	75.44	76.62
	GPT-4o	89.10	86.65	77.10	81.58	83.61
	DoubleCast	90.70	84.45	71.20	61.99	77.09
	MoiraiAgent	91.35	86.20	79.80	67.25	81.15
$1\times$	Mistral (7B)	71.80	68.10	54.30	70.66	66.21
	Llama-3 (70B)	80.90	77.60	65.21	79.53	75.81
	GPT-4o	87.79	84.61	73.90	85.09	82.85
	DoubleCast	89.40	82.10	67.80	66.47	76.44
	MoiraiAgent	90.11	83.90	76.51	71.83	80.59
$2\times$	Mistral (7B)	68.80	64.65	49.25	76.90	64.90
	Llama-3 (70B)	78.10	73.95	59.80	84.80	74.16
	GPT-4o	85.45	81.05	68.80	89.77	81.27
	DoubleCast	87.10	78.00	62.15	72.81	75.02
	MoiraiAgent	87.90	79.80	71.30	77.19	79.05

Table 24: Sign-only sensitivity of relational criteria, with $\eta_D = 0$.

Model	Direction	Contrast	Intensity
Mistral (7B)	75.62	73.18	61.84
Llama-3 (70B)	84.06	82.17	72.48
GPT-4o	90.53	88.61	81.24
DoubleCast	92.01	86.82	75.93
MoiraiAgent	92.64	88.47	83.52

We define

$$\epsilon_D = \rho v_D, \qquad \rho = 0.1, \tag{32}$$

using the same pre-specified coefficient for every domain. The current release uses the same scale for both practical thresholds,

$$\eta_D = \tau_D = \epsilon_D. \tag{33}$$

Applying this rule gives $\eta_D = \tau_D = 0.01$, 0.02, and 0.10 for Energy, Economy, and Politics, respectively. If no non-zero change is available, the documented numerical resolution of the target is used as the fallback scale.

D.3.3 Tolerance Sensitivity

We repeat Relational Comparison with $\rho \in \{0.05, 0.10, 0.20\}$, corresponding to $0.5\times$, $1\times$, and $2\times$ the default tolerance. The same multiplier is applied to both η_D and τ_D in every domain.

As expected, larger tolerances make Direction, Contrast, and Intensity stricter and Restraint more permissive. Despite these shifts, the category-level response profiles remain similar, and Overall SLA remains relatively stable across the tested settings.

Sign-only sensitivity. We additionally evaluate Direction, Contrast, and Intensity with $\eta_D = 0$, removing the practical magnitude requirement. Restraint is omitted because it is inherently defined by revision magnitude.

The relative response profiles remain similar under sign-only evaluation, indicating that the reported relational patterns are not driven by the particular practical tolerance.

D.4 Grounded Comparison Details

Recall that $\Delta_{i,h}(s) = \widehat{y}^{s}_{i,h} - \widehat{y}^{0}_{i,h}$ denotes the scenario-induced revision, and $g_i(s)$ its mean squared discrepancy from the neutral-relative deviations of the empirical matched cases.

D.4.1 Empirical References for Evaluation

Grounded Comparison uses the fixed empirical matched set $\mathcal{P}_i(s)$ for each supported anchor–scenario pair (i, s). For matched case j, the evaluated model generates its own neutral forecast and defines $d_{j,h} = y_{j,h} - \widehat{y}^{0}_{j,h}$. Thus, both anchor revisions and matched-case deviations are measured relative to the evaluated model's own neutral baseline; the construction-reference forecasts used for matching are not used as evaluation references.

D.4.2 Pair and Matched-Case Weighting

Forecast steps and matched cases receive equal weight within each pair, and supported pairs receive equal weight within each domain. A real case reused across multiple empirical sets contributes independently to each corresponding pair. GSG is macro-averaged with equal domain weight, while GCE is reported separately by domain.

D.4.3 No-Revision Baseline

The no-revision baseline uses the same pairs, matched sets, and weighting, but sets $\Delta_{i,h}(s) = 0$. GSG therefore measures improvement in empirical alignment relative to retaining the neutral forecast.

D.4.4 Interpretation of GCE and GSG

Matched real cases are empirical references rather than ground truth for the unobserved alternative future. GCE is an RMSE-like discrepancy in the target unit: lower values indicate closer alignment, but zero does not represent a uniquely correct alternative future. GSG indicates whether scenario conditioning improves this alignment relative to making no revision.

E Experimental Setup

E.1 Evaluated Models

We evaluate three groups of models: directly prompted LLMs, context-aware forecasting systems, and time-series foundation models (TSFMs).

The directly prompted LLMs include Mistral-7B-Instruct v0.3 (Jiang et al., 2023), Llama-3-70B-Instruct (Grattafiori et al., 2024), and GPT-4o (OpenAI, 2024).

For context-aware forecasting, we evaluate DoubleCast (Zheng et al., 2026) and MoiraiAgent (Yan et al., 2026). For DoubleCast, we retain the released architecture and Chronos-T5-Large numerical backbone, replace its original Qwen3-14B text encoder with Qwen1.5-14B-Chat, and perform cross-modal alignment on CAF-7M before evaluation, as detailed below. For MoiraiAgent, we follow the released GIFT-CTX implementation, which uses Chronos-2 as its default time-series forecasting tool, and use GPT-4o as its LLM component. Agentic frameworks such as Nexus (Das et al., 2026) can incorporate explicit relevance-assessment steps, but their task-specific agent configuration and prompting do not fit the fixed inference interface used for the comparison. We therefore exclude them from the main experimental comparison.

The numerical TSFM baselines include Chronos-2 (Ansari et al., 2025), TimesFM-3 (Jain &Sen, 2026), and TiRex-2 (Podest et al., 2026).

E.2 Model Versions and Knowledge Cutoffs

Table 25 reports the forecasting and textual components used in our experiments. For LLM-based components, we restrict model selection to models with explicitly documented knowledge cutoffs

Table 25: Model components and documented LLM knowledge cutoffs. A dash indicates that no textual LLM is used.

System	Time-series component	Text / LLM component	Knowledge cutoff
Mistral (7B)	–	Mistral-7B-Instruct v0.3	Sep. 2023
Llama-3 (70B)	–	Llama-3-70B-Instruct	Dec. 2023
GPT-4o	–	GPT-4o	Oct. 2023
DoubleCast	Chronos-T5-Large	Qwen1.5-14B-Chat	Sep. 2023
MoiraiAgent	Chronos-2	GPT-4o	Oct. 2023
Chronos-2	amazon/chronos-2	–	–
TimesFM-3	google/timesfm-3.0-pytorch	–	–
TiRex-2	NX-AI/TiRex-2	–	–

that precede the WIT evaluation period. Because textual knowledge cutoffs do not characterize numerical time-series pretraining, possible overlap with WIT is audited separately below.

DoubleCast text encoder. The released DoubleCast configuration uses Chronos-T5-Large as its numerical backbone and Qwen3-14B as its text encoder. Because Qwen3-14B does not satisfy our requirement for an explicitly documented pre-WIT knowledge cutoff, we replace it with Qwen1.5-14B-Chat, whose documented knowledge cutoff is September 2023. Qwen1.5-14B-Chat retains the 5,120-dimensional hidden representation expected by the released DoubleCast cross-attention interface, allowing the model architecture to remain unchanged. Because replacing the encoder changes its semantic representation despite this dimensional compatibility, we re-align the text–time-series fusion components on CAF-7M before evaluation. This alignment uses the same training source as DoubleCast and does not use any WIT numerical series, scenarios, factual evidence, or empirical matched cases. We refer to this aligned configuration as DoubleCast throughout the experiments.

Time-series pretraining overlap audit. We separately inspect the disclosed pretraining sources of all numerical forecasting components, since LLM knowledge cutoffs do not address possible time-series contamination. The Chronos-T5 backbone used by DoubleCast, the Chronos-2 forecasting tool used by MoiraiAgent, and the Chronos-2, TimesFM-3, and TiRex-2 numerical baselines were pretrained on established time-series corpora, with synthetic or augmented time series additionally used by the respective methods. We verified the disclosed training sources against WIT and found no overlap with the Energy, Economy, or Politics evaluation data used in the benchmark. Accordingly, none of the time-series forecasting components is trained on the numerical data used for WIT evaluation.

E.3 Inference and Implementation Settings

The directly prompted GPT-4o baseline and the GPT-4o component of MoiraiAgent are evaluated through API inference. Mistral-7B-Instruct and Llama-3-70B-Instruct are served locally using vLLM on NVIDIA RTX PRO 6000 Blackwell GPUs with 96 GB of GPU memory. DoubleCast uses the released implementation with the CAF-7M-aligned configuration described above, while the numerical TSFMs use their released inference implementations and checkpoints.

For generative LLM inference, we use temperature 1.0 and top-$p = 1.0$ where these parameters are exposed by the inference interface, corresponding to no additional temperature scaling or nucleus truncation. Frequency and presence penalties are set to zero where applicable. The maximum output length is 512 tokens, which is sufficient for the required 20-step numerical trajectory. Decoding settings are held fixed across domains and scenario conditions.

Directly prompted LLM outputs must parse as a JSON array containing exactly $H = 20$ finite numerical values. If parsing fails or the output length is incorrect, the request is retried with the same forecasting inputs and a formatting-only reminder, for at most two retries after the initial attempt. No numerical values are manually corrected or imputed.

Table 26: Forecast errors of numerical-only time-series foundation models on the WIT factual evaluation anchors. All models receive only the numerical history and report the median forecast ($q = 0.5$). Errors are reported separately by domain because target scales differ. Lower is better.

Model	Energy	Economy	Politics
Chronos-2	0.269	2.272	23.340
TimesFM-3	**0.256**	2.222	24.051
TiRex-2	0.260	**2.177**	**22.875**

For Chronos-2, TimesFM-3, and TiRex-2, we use the median forecast ($q = 0.5$) with their released inference configurations. These models receive only the numerical history and serve as numerical-only baselines for factual forecasting performance.

E.4 Repeated-Run Protocol

Each scenario-conditioned system is evaluated over three runs using identical benchmark inputs and inference settings. Where deterministic seed control is available, we use seeds 42, 43, and 44 for the three runs. For API components without reliable deterministic seed control, the runs consist of independent calls under otherwise identical settings.

Within each run, one neutral forecast is generated per anchor and reused across all scenario comparisons associated with that anchor. For Grounded Comparison, one neutral forecast is likewise generated for each matched real case and reused wherever that case contributes within the run. Metrics are computed independently for each run and reported as the mean and sample standard deviation across the three runs.

The numerical TSFM baselines use a single median forecast per input and are evaluated once.

F Additional Results and Analyses

F.1 Numerical-Only Forecasting Baselines

Table 26 reports domain-wise errors for the numerical-only TSFM baselines. These models receive only the numerical history and therefore provide a reference for the predictive information available without textual context. Compared with the neutral forecasts of the scenario-conditioned systems, the results show no systematic advantage from providing the fixed domain background alone. This is consistent with its intended role: $\mathcal{C}_D$ specifies general target mechanisms and applicability conditions but excludes anchor-specific events, dated outlooks, and other information that would directly reveal the future trajectory.

F.2 Domain-Wise Factual Comparison

Table 27 provides the full domain-wise Factual Comparison. Factual scenarios improve average forecast accuracy for every model and domain, but the magnitude of the gain varies substantially. GPT-4o shows the largest relative gains in Energy and Economy, whereas the context-aware systems benefit more strongly in Politics. CWR follows a different pattern, showing that larger average gains do not necessarily correspond to improvements across a larger share of anchors.

F.3 Domain-Wise Relational Comparison

Tables 28–30 report the four relational categories separately for each domain. Across domains, models generally perform better on Direction and Contrast than on Intensity and Restraint, indicating that coarse scenario responses are easier than calibrating response strength or ignoring irrelevant information. The category profiles also differ across models, so similar Overall SLA can hide substantially different strengths and failure modes.

Table 27: Full domain-wise factual scenario response results on WIT. Factual Scenario Gain (FSG) and Context Win Rate (CWR) are reported for each domain; Overall denotes the macro-average across Energy, Economy, and Politics. Higher is better for both metrics.

Model	Factual Scenario Gain (%)				Context Win Rate (%)			
	Energy	Economy	Politics	Overall	Energy	Economy	Politics	Overall
Directly prompted LLMs								
Mistral (7B)	7.42 ±0.58	7.45 ±0.66	6.71 ±0.65	7.19 ±0.63	57.02 ±1.52	55.26 ±2.63	54.39 ±1.52	55.56 ±1.83
Llama-3 (70B)	11.82 ±0.64	11.58 ±0.72	10.35 ±0.71	11.25 ±0.69	63.16 ±2.63	60.53 ±2.63	58.77 ±1.52	60.82 ±2.21
GPT-4o	**14.77 ±0.56**	**13.45 ±0.65**	16.54 ±0.71	**14.92 ±0.64**	66.67 ±1.52	65.79 ±2.63	64.04 ±1.52	65.50 ±1.83
Context-aware forecasting systems								
DoubleCast	10.58 ±0.36	11.09 ±0.40	18.47 ±0.50	13.38 ±0.42	66.67 ±1.52	**67.54 ±1.52**	64.04 ±1.52	66.08 ±1.34
MoiraiAgent	11.03 ±0.67	9.61 ±0.78	**19.19 ±0.92**	13.28 ±0.79	**68.42 ±2.63**	65.79 ±2.63	**67.54 ±1.52**	**67.25 ±2.21**

F.4 Domain-Wise Grounded Comparison

Table 31 reports Grounded Comparison separately by domain. Across all three domains, the context-aware forecasting systems show stronger empirical alignment than the directly prompted LLMs, although the magnitude of the advantage varies by domain. The results also reinforce that relational consistency and alignment with neutral-relative deviations observed in matched real cases capture distinct aspects of scenario responsiveness.

F.5 Anchor-Level Analysis

We further separate anchors according to whether the factual scenario improves forecast error relative to the neutral condition. As shown in Tables 32–34, anchors with factual improvements consistently exhibit lower relational violation rates, although substantial violations remain within the improved subset. Factual gains are therefore associated with stronger relational consistency but do not ensure appropriate responses across scenarios.

G Limitations and Future Work

WIT evaluates a selected set of challenging cases rather than all real-world forecasting settings. The current release focuses on carefully verified anchors across three domains, with screening designed to identify informative scenario-conditioned forecasting cases. Future work can extend the same construction protocol to larger collections, additional domains and temporal regimes, and alternative reference ensembles to assess how broadly the observed behaviors generalize.

Grounded Comparison uses matched real cases as empirical references rather than counterfactual ground truth. Its current coverage is also limited by the requirement for both numerical and semantic compatibility between matched cases. As larger case libraries become available, future versions can support more matched cases per scenario, richer empirical response ranges, and probabilistic or uncertainty-aware evaluation.

The current relational evaluation focuses on Direction, Contrast, Intensity, and Restraint under controlled scenario changes. More complex settings, including interacting events, ambiguous scenarios, and richer temporal effects, remain outside the current scope. Future work can expand the range of scenario types and response relations while retaining the same fixed-state comparison principle, enabling WIT to grow without changing its core evaluation framework.

Table 28: Full Relational Comparison results for Energy. Scenario Logic Accuracy (SLA) is reported for Direction, Contrast, Intensity, and Restraint; Overall is their equal-weight average. Higher is better.

Model	Direction	Contrast	Intensity	Restraint	Overall
Directly prompted LLMs					
Mistral (7B)	73.27±1.54	69.36±1.41	55.79±2.11	72.22±2.21	67.66±1.57
Llama-3 (70B)	82.11±1.10	78.71±1.23	66.58±1.71	80.70±1.75	77.02±1.27
GPT-4o	88.54±0.75	**85.26±0.80**	74.87±1.45	**85.96±1.75**	**83.66±1.06**
Context-aware forecasting systems					
DoubleCast	89.88±0.70	82.34±0.79	68.20±1.25	66.96±1.34	76.85±1.02
MoiraiAgent	**90.70±0.96**	84.09±0.88	**77.24±1.45**	72.51±1.34	81.14±1.04

Table 29: Full Relational Comparison results for Economy. Scenario Logic Accuracy (SLA) is reported for Direction, Contrast, Intensity, and Restraint; Overall is their equal-weight average. Higher is better.

Model	Direction	Contrast	Intensity	Restraint	Overall
Directly prompted LLMs					
Mistral (7B)	72.00±1.59	68.42±1.54	54.71±2.06	70.76±2.21	66.47±1.69
Llama-3 (70B)	81.10±1.04	77.78±0.99	65.46±1.64	79.82±1.75	76.04±1.21
GPT-4o	87.94±0.77	84.43±0.88	74.12±1.23	**85.38±1.34**	**82.97±0.93**
Context-aware forecasting systems					
DoubleCast	**90.31±0.71**	82.82±0.77	68.59±1.15	66.67±1.75	77.10±1.10
MoiraiAgent	90.20±0.82	**84.72±0.99**	**77.08±1.23**	71.93±1.75	80.99±1.11

Table 30: Full Relational Comparison results for Politics. Scenario Logic Accuracy (SLA) is reported for Direction, Contrast, Intensity, and Restraint; Overall is their equal-weight average. Higher is better.

Model	Direction	Contrast	Intensity	Restraint	Overall
Directly prompted LLMs					
Mistral (7B)	70.11±1.51	66.52±1.43	52.39±2.12	69.01±2.21	64.51±1.64
Llama-3 (70B)	79.50±1.02	76.32±1.10	63.60±1.61	78.07±1.75	74.37±1.21
GPT-4o	86.91±0.73	**84.14±0.89**	72.71±1.39	**83.92±1.34**	**81.92±0.96**
Context-aware forecasting systems					
DoubleCast	88.01±0.68	81.14±0.88	66.62±1.24	65.79±1.75	75.39±1.14
MoiraiAgent	**89.41±0.93**	82.89±1.10	**75.19±1.39**	71.05±1.75	79.64±1.20

Table 31: Full domain-wise Grounded Scenario Gain (GSG) results on WIT. Higher is better. Overall denotes the equal-domain macro-average across Energy, Economy, and Politics, regardless of the number of supported anchor–scenario pairs.

Model	Energy (3 supported pairs)	Economy (11 supported pairs)	Politics (6 supported pairs)	Overall
Directly prompted LLMs				
Mistral (7B)	−4.18±1.20	−2.80±0.85	−6.70±1.45	−4.56±1.17
Llama-3 (70B)	9.60±0.55	8.20±0.42	6.80±0.63	8.20±0.50
GPT-4o	12.53±1.05	11.40±0.70	10.20±0.90	11.38±0.86
Context-aware forecasting systems				
DoubleCast	**17.83±0.48**	**18.60±0.34**	16.50±0.55	**17.64±0.43**
MoiraiAgent	17.39±0.72	18.00±0.52	**17.10±0.78**	17.50±0.64

Table 32: Anchor-level analysis for Energy over 38 anchors and three runs. Improved anchors satisfy $e_i^{s^*} < e_i^0$; ties belong to the not-improved subset. Direction, Contrast, and Intensity use controlled nonfactual scenarios, excluding the factual scenario. Restraint violations indicate excessive revision under irrelevant scenarios. Violation rates (%, lower is better) are averaged within anchors and then across each subset. Anchor counts follow run order 0/1/2.

Model	Factual outcome	Anchors per run	Direction	Contrast	Intensity	Restraint
Mistral (7B)	Improved	21/22/22	$22.87_{\pm 1.54}$	$26.33_{\pm 1.56}$	$39.50_{\pm 2.24}$	$22.05_{\pm 2.28}$
	Not improved	17/16/16	$31.84_{\pm 1.58}$	$36.36_{\pm 1.51}$	$50.45_{\pm 2.29}$	$35.38_{\pm 2.08}$
Llama-3 (70B)	Improved	23/25/24	$15.30_{\pm 1.20}$	$18.31_{\pm 1.35}$	$29.74_{\pm 1.96}$	$15.76_{\pm 1.76}$
	Not improved	15/13/14	$22.35_{\pm 1.31}$	$26.39_{\pm 1.39}$	$39.73_{\pm 1.92}$	$25.37_{\pm 1.45}$
GPT-4o	Improved	25/26/25	$9.81_{\pm 0.79}$	$12.71_{\pm 0.81}$	$22.48_{\pm 1.52}$	$11.40_{\pm 2.00}$
	Not improved	13/12/13	$14.76_{\pm 0.86}$	$18.80_{\pm 1.07}$	$30.44_{\pm 1.62}$	$19.30_{\pm 1.29}$
DoubleCast	Improved	25/25/26	$8.45_{\pm 0.75}$	$15.34_{\pm 0.88}$	$28.80_{\pm 1.40}$	$28.50_{\pm 1.59}$
	Not improved	13/13/12	$13.44_{\pm 0.81}$	$22.31_{\pm 0.92}$	$37.80_{\pm 1.32}$	$42.17_{\pm 1.97}$
MoiraiAgent	Improved	26/25/27	$8.04_{\pm 0.95}$	$14.00_{\pm 0.90}$	$20.55_{\pm 1.60}$	$24.33_{\pm 1.45}$
	Not improved	12/13/11	$12.02_{\pm 0.86}$	$20.03_{\pm 1.07}$	$27.55_{\pm 1.44}$	$34.34_{\pm 1.75}$

Table 33: Anchor-level analysis for Economy over 38 anchors and three runs. Definitions and aggregation follow Table 32.

Model	Factual outcome	Anchors per run	Direction	Contrast	Intensity	Restraint
Mistral (7B)	Improved	20/21/22	$24.00_{\pm 1.60}$	$27.11_{\pm 1.65}$	$40.34_{\pm 2.13}$	$23.31_{\pm 2.09}$
	Not improved	18/17/16	$32.94_{\pm 1.64}$	$37.12_{\pm 1.54}$	$51.40_{\pm 2.34}$	$36.61_{\pm 2.26}$
Llama-3 (70B)	Improved	22/24/23	$16.15_{\pm 1.24}$	$19.05_{\pm 1.18}$	$30.58_{\pm 1.72}$	$16.45_{\pm 1.85}$
	Not improved	16/14/15	$23.11_{\pm 1.17}$	$27.09_{\pm 1.32}$	$40.60_{\pm 1.91}$	$25.91_{\pm 1.34}$
GPT-4o	Improved	24/26/25	$10.37_{\pm 0.87}$	$13.54_{\pm 0.94}$	$23.13_{\pm 1.42}$	$12.00_{\pm 1.16}$
	Not improved	14/12/13	$15.31_{\pm 0.87}$	$19.47_{\pm 1.03}$	$31.15_{\pm 1.44}$	$19.61_{\pm 1.75}$
DoubleCast	Improved	26/25/26	$8.05_{\pm 0.74}$	$14.92_{\pm 0.86}$	$28.48_{\pm 1.19}$	$28.97_{\pm 2.10}$
	Not improved	12/13/12	$13.09_{\pm 0.86}$	$21.87_{\pm 0.93}$	$37.53_{\pm 1.35}$	$42.38_{\pm 1.82}$
MoiraiAgent	Improved	25/24/26	$8.42_{\pm 0.90}$	$13.23_{\pm 0.80}$	$20.55_{\pm 1.35}$	$24.49_{\pm 1.71}$
	Not improved	13/14/12	$12.44_{\pm 0.96}$	$19.21_{\pm 1.16}$	$27.48_{\pm 1.49}$	$34.98_{\pm 1.43}$

Table 34: Anchor-level analysis for Politics over 38 anchors and three runs. Definitions and aggregation follow Table 32.

Model	Factual outcome	Anchors per run	Direction	Contrast	Intensity	Restraint
Mistral (7B)	Improved	20/21/21	$25.78_{\pm 1.57}$	$29.01_{\pm 1.53}$	$42.61_{\pm 2.23}$	$24.76_{\pm 2.29}$
	Not improved	18/17/17	$34.79_{\pm 1.66}$	$38.82_{\pm 1.62}$	$53.57_{\pm 2.32}$	$38.42_{\pm 2.01}$
Llama-3 (70B)	Improved	22/23/22	$17.62_{\pm 1.09}$	$20.38_{\pm 1.03}$	$32.31_{\pm 1.75}$	$17.94_{\pm 1.89}$
	Not improved	16/15/16	$24.61_{\pm 1.19}$	$28.40_{\pm 1.48}$	$42.24_{\pm 1.74}$	$27.64_{\pm 1.34}$
GPT-4o	Improved	24/25/24	$11.30_{\pm 0.74}$	$13.70_{\pm 0.76}$	$24.41_{\pm 1.43}$	$13.22_{\pm 1.25}$
	Not improved	14/13/14	$16.29_{\pm 0.96}$	$19.70_{\pm 0.98}$	$32.43_{\pm 1.64}$	$21.18_{\pm 2.03}$
DoubleCast	Improved	24/24/25	$10.19_{\pm 0.68}$	$16.31_{\pm 1.03}$	$30.19_{\pm 1.39}$	$29.24_{\pm 1.28}$
	Not improved	14/14/13	$15.20_{\pm 0.91}$	$23.40_{\pm 0.88}$	$39.06_{\pm 1.35}$	$43.04_{\pm 2.11}$
MoiraiAgent	Improved	26/25/26	$9.28_{\pm 0.93}$	$15.15_{\pm 1.13}$	$22.49_{\pm 1.53}$	$25.56_{\pm 1.16}$
	Not improved	12/13/12	$13.31_{\pm 0.93}$	$21.17_{\pm 1.04}$	$29.62_{\pm 1.41}$	$35.97_{\pm 2.57}$