AUDITING INFERENCE-TIME ABSTENTION GATES IN LLM TRADING AGENTS WITH FIXED-OUTCOME REPLAY

Anonymous authors
Paper under double-blind review

ABSTRACT

Agentic large language models (LLMs) increasingly use inference-time abstention and fallback mechanisms, but aggregate utility does not reveal which decisions those mechanisms change. We study this attribution problem with Fixed-Outcome Replay (FOR), a diagnostic protocol that logs each proposal, gate trigger, deployed action, and realized outcome, and separates end-to-end system contrasts from proposal-fixed controller contrasts. On the canonical 3,214-decision MAG7 corpus spanning January 2024–October 2025, the ungated A5 baseline exceeds the A1 full system by 17.19 percentage points (+23.56% versus +6.37%), but this is an end-to-end contrast. Holding the A1 proposals fixed, the default controller changes realized utility by -1.81 percentage points (95% bootstrap interval [-6.40, 2.15]), which is inconclusive. Trigger-level replay gives +3.70 percentage points for low-confidence interventions and -5.48 percentage points for divergence-only interventions. FOR is therefore an accounting diagnostic for proposal coverage, trigger composition, conditional risk, and logged rejected-action value under fixed outcomes; it does not estimate deployment-time causal effects.

1 INTRODUCTION

Agentic LLMs can synthesize heterogeneous information into coherent narratives, but persuasive explanations do not imply calibrated decisions. This mismatch is especially consequential in noisy settings such as financial markets, where an apparently well-supported action can remain poorly grounded in the realized outcome. Studies of verbalized uncertainty and in-context calibration document this gap (Groot & Valdenegro-Toro, 2024; Li et al., 2025a; Zhang et al., 2024; Ulmer et al., 2024; Kumar et al., 2024; Epstein et al., 2025). Calibration can also change after alignment or depend on internal belief signals rather than verbal scores (Xiao et al., 2025; 2026; Zeng et al., 2025). Common responses include self-reflection (Shinn et al., 2023), cross-model guidance (Wang et al., 2024; Zhang et al., 2025a), and selective abstention (Feng et al., 2024; Kirichenko et al., 2025; Wen et al., 2025; Madhusudhan et al., 2025; Nguyen et al., 2026; Phillips et al., 2026).

Recent financial LLM systems, including TradingAgents (Xiao et al., 2024), FINCON (Yu et al., 2024), StockBench (Chen et al., 2025), and TS-Agent (Ang et al., 2025), commonly report end-to-end outcomes. Related finance-oriented workflows include FinGPT, interactive co-reasoning, and cross-view correction for time-series forecasting (Yang et al., 2023; Zhang et al., 2025b; Li et al., 2025b); empirical TSFM studies further emphasize domain shift and finance-specific adaptation (Rahimikia et al., 2025). Such outcomes are useful for system description, but they do not identify whether a change came from semantic proposal generation, an abstention gate, or a different exposure pattern. The diagnostic question is therefore narrower: when a controller overrides a proposal, what realized value is associated with that rejected action?

We address this attribution gap by writing the controller as a selective decision rule and by treating the time-series foundation model as an operational signal, not as ground truth. Chronos, TimesFM, Moirai, Lag-Llama, Moirai-MoE, and Moirai 2.0 provide probabilistic forecasting interfaces (Ansari et al., 2024; Das et al., 2024; Woo et al., 2024; Rasul et al., 2024; Liu et al., 2025b;a); recent work also studies foundation-model forecasts inside conformal prediction (Achour et al., 2025; Chen et al., 2024). Our experiments use a Chronos-family forecast model. The model supplies an operational

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.

confidence and directional-probability signal, but its calibration and economic usefulness remain empirical questions. The noisy financial environment and market microstructure motivate this caution (Brogaard et al., 2014).

We instrument a role-specialized multi-agent system with a deterministic controller and introduce a matched replay protocol. The protocol logs the proposal, forecast-model variables, trigger identity, deployed action, and next-day payoff for each decision. Under the stated fixed-outcome assumption, the resulting identity gives an accounting decomposition; it is not a general causal or off-policy estimator. The canonical corpus contains 3,214 MAG7 stock-day decisions per variant over January 2024–October 2025. A retained 140-row slice is kept separately as a local audit artifact and does not populate the canonical result tables.

The reported results make the estimand distinction visible. A5 obtains +23.56% and A1 obtains +6.37%, an end-to-end difference of +17.19 percentage points that combines proposal, participation, and controller changes. Holding the A1 proposals fixed, the default controller changes utility from +8.18% to +6.37%, a difference of -1.81 percentage points with interval [-6.40, 2.15]. The separate A1-GR fixed-proposal raw-divergence replay changes A1 by -3.52 percentage points with interval [-8.90, 0.75]. At the default operating point, A1 deploys nonzero actions on 46.92% of decisions with 41.91% conditional directional risk. These results motivate risk–coverage and trigger-specific reporting rather than ranking abstention layers by aggregate return alone.

Our contribution is an audit protocol rather than a deployable trading algorithm or a new forecasting model. The financial domain is used as a noisy testbed for explicit outcome and exposure accounting. Our contributions are:

- **Estimand separation:** We distinguish the end-to-end system contrast from proposal-fixed gate and trigger contrasts, and state the outcome-invariance assumption needed for the replay identity.
- **Instrumented replay:** We specify the telemetry needed to reproduce proposal coverage, fallback composition, deployed coverage, conditional directional risk, and logged rejected-action value on the canonical 3,214-decision log.
- **Empirical diagnosis:** We report end-to-end, proposal-fixed, and trigger-level quantities together with block-bootstrap intervals and risk–coverage diagnostics, showing why aggregate return alone is insufficient to interpret an inference-time controller.

2 Related Work

Agentic workflows and evaluation paradigms. Agentic financial systems such as TradingAgents, FinCon, StockBench, and TS-Agent study multi-agent reasoning, realistic market inputs, and time-series workflows (Xiao et al., 2024; Yu et al., 2024; Chen et al., 2025; Ang et al., 2025). These papers motivate the testbed, but their end-to-end results do not by themselves identify the contribution of a downstream abstention gate. We therefore focus on the intervention estimand and log the rejected actions explicitly.

Selective prediction and safety abstention. Calibration and abstention work studies when a model should defer or decline, including verbalized uncertainty, knowledge-gap detection, and benchmarked unanswerability (Groot & Valdenegro-Toro, 2024; Li et al., 2025a; Feng et al., 2024; Kirichenko et al., 2025; Wen et al., 2025; Madhusudhan et al., 2025; Nguyen et al., 2026; Phillips et al., 2026). Selective-classification work makes the risk–coverage trade-off explicit (Geifman & El-Yaniv, 2017; 2019). Learning-to-defer formalizes the decision boundary between a model and an expert (Mozannar & Sontag, 2020). Our controller does not train a selective classifier: it instantiates a logged hold action and evaluates the conditional risk and logged value of the decisions it rejects.

Inference-time control and forecast signals. Inference-time alignment methods steer outputs without retraining, including cross-model guidance and controllable safety adaptation (Wang et al., 2024; Zhang et al., 2025a; Banerjee et al., 2025; Bhattacharjee et al., 2024; Hazra et al., 2024). Probabilistic time-series models provide a separate forecasting interface, but a predictive distribution is not automatically a calibrated safety certificate (Ansari et al., 2024; Das et al., 2024; Woo et al., 2024; Rasul et al., 2024; Liu et al., 2025b;a). Off-policy evaluation and conformal risk control make the identification requirements explicit (Jiang & Li, 2016; Angelopoulos et al., 2024; Achour et al.,

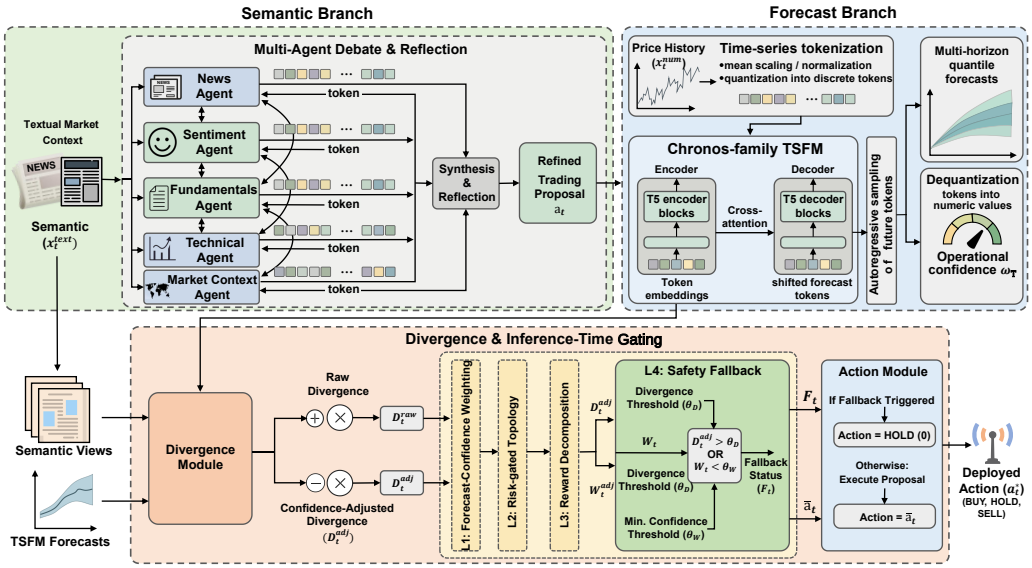

Figure 1: Instrumented diagnostic testbed architecture. The semantic branch produces a debated and reflected proposal. The forecast branch provides a predictive distribution and an operational confidence score. The diagnostic analysis explicitly isolates confidence-weighted disagreement and fallback to a hold state, while other implementation controls remain fixed across matched evaluations.

2025; Chen et al., 2024). Our replay protocol is deliberately narrower: it is an accounting diagnostic under fixed logged outcomes, not a substitute for randomized logging or a validated off-policy estimator.

3 Fixed-Outcome Replay Framework

3.1 Instrumented diagnostic architecture

Figure 1 summarizes the diagnostic testbed. The semantic branch contains five large language model (LLM) analyst agents responsible for news, sentiment, fundamentals, technical analysis, and broader market context. Their outputs are consolidated through an iterative debate and reflection mechanism, drawing on verbal self-correction (Shinn et al., 2023). Concurrently, the forecast branch uses a Chronos-family time-series foundation model (TSFM) to generate quantile forecasts and an operational confidence score (Ansari et al., 2024). A divergence module compares the semantic and forecast views, after which the deterministic controller either executes the proposal or maps it to a neutral hold state.

The controller implementation encompasses four inference-time components. Confidence weighting scales normalized semantic–forecast disagreement by the forecast model's operational confidence. Risk-gated topology and reward decomposition are fixed implementation controls in the present comparison. Abstention fallback is the final selective decision rule. Because the evaluation is based on logged matched decisions, the diagnostic focuses on replayable contrasts for confidence weighting and fallback; the other controls are not treated as isolated intervention variables.

3.2 Selective decision formulation

Let $t \in \{1, \ldots, n\}$ index a stock-day decision. The state is $x_t = (x_t^{\text{text}}, x_t^{\text{num}})$, where the two components contain the chronologically available textual and numerical context. The semantic branch proposes a pre-fallback action $\bar{a}_t \in \mathcal{A} = \{-1, 0, +1\}$, with $+1$ denoting positive exposure, 0 a natural hold, and -1 negative exposure. The forecast model returns a predictive distribution Q_t for a ten-trading-day horizon and an operational confidence score $w_t \in [0, 1]$. Define the semantic

directional probability and the forecast directional probability by

$$p_t^{\mathrm{sem}} = \frac{\bar{a}_t + 1}{2}, \qquad p_t^{\mathrm{tsfm}} = \mathbb{P}_{Y \sim Q_t}(Y > 0), \tag{1}$$

and define the unnormalized disagreement

$$G_t = \left| p_t^{\mathrm{sem}} - p_t^{\mathrm{tsfm}} \right|. \tag{2}$$

The protocol fixes $G_{\max} = 0.30$ from a 60-trading-day antecedent validation segment and freezes it before each evaluated hold-out block. The candidate thresholds are $\theta_D \in \{0.15, 0.20, 0.25, 0.30, 0.40, 0.50\}$ and $\theta_w \in \{0.10, 0.15, 0.20, 0.22, 0.25, 0.30\}$. The selection rule minimizes validation directional error subject to deployed nonzero coverage in [45%,65%], with validation utility used only as a tie-breaker; no holdout return enters selection. The six-by-six sweep is a finite sensitivity grid, not a fitted calibration guarantee. The normalized disagreement is

$$D_t^{\mathrm{norm}} = \min(G_t / G_{\max}, 1) \tag{3}$$

and the confidence-adjusted disagreement is

$$D_t^{\mathrm{adj}} = w_t D_t^{\mathrm{norm}}. \tag{4}$$

The default gate uses $D_t^{\mathrm{gate}} = D_t^{\mathrm{adj}}$. In the A1-GR controller-only replay, the gate uses $D_t^{\mathrm{gate}} = D_t^{\mathrm{norm}}$, while the low-confidence safeguard remains active. Thus the fallback event is the union

$$F_t(\theta_D, \theta_w) = \{D_t^{\mathrm{gate}} > \theta_D\} \cup \{w_t < \theta_w\}, \tag{5}$$

with default thresholds $\theta_D = 0.30$ and $\theta_w = 0.22$. The deployed action is

$$a_t^* = \bar{a}_t \, \mathbb{I}\{F_t^c\}. \tag{6}$$

The fallback maps a rejected proposal to a hold. The TSFM horizon and payoff horizon are distinct: Q_t is produced for ten trading days, whereas the primary realized payoff $y_t = y_t^{(1)}$ is the next-trading-day close-to-close return. We use signed utility $U(a, y) = ay$ and the directional error loss $\ell(a, y) = \mathbb{I}\{a \neq 0,\ ay < 0\}$. The latter assigns zero error to a hold by construction and is used only as a conditional diagnostic.

3.3 Deterministic controller identities

The following results describe deterministic properties of the controller and the accounting estimand used in replay. They are audit identities, not a claim of deep theoretical novelty, threshold optimality, calibrated risk guarantees, or a general causal effect. The inclusion and monotonicity propositions are properties of the logged scores; the payoff identity additionally requires an outcome-invariance assumption.

Definition 1 (Trigger sets)**.** For a divergence threshold $\theta_D > 0$, define

$$\mathcal{T}_{\mathrm{norm}}(\theta_D) = \{t : D_t^{\mathrm{norm}} > \theta_D\}, \qquad \mathcal{T}_{\mathrm{adj}}(\theta_D) = \{t : w_t D_t^{\mathrm{norm}} > \theta_D\}.$$

Proposition 1 (Confidence weighting can only remove divergence triggers)**.** If $w_t \in [0, 1]$ for all t, then $\mathcal{T}_{\mathrm{adj}}(\theta_D) \subseteq \mathcal{T}_{\mathrm{norm}}(\theta_D)$. Moreover, a point t is removed by confidence weighting exactly when

$$D_t^{\mathrm{norm}} > \theta_D \quad \text{and} \quad w_t \leq \frac{\theta_D}{D_t^{\mathrm{norm}}}.$$

Proposition 1 states the set relationship induced by multiplying a score by $w_t \in [0, 1]$. It does not say that the resulting confidence score is calibrated or that every removed trigger is harmful.

Proposition 2 (Threshold monotonicity)**.** Let $F_t(\theta_D, \theta_w)$ be defined as in Eq. (5). If $\theta'_D \geq \theta_D$ and $\theta'_w \leq \theta_w$, then $F_t(\theta'_D, \theta'_w) \subseteq F_t(\theta_D, \theta_w)$. Consequently, the participation rate

$$C(\theta_D, \theta_w) = \frac{1}{n} \sum_{t=1}^{n} \mathbb{I}\{F_t(\theta_D, \theta_w)^c, \bar{a}_t \neq 0\}$$

is nondecreasing in θ_D and nonincreasing in θ_w.

This proposition describes the controller's participation frontier for fixed logged scores. It does not imply that a less conservative setting has better realized utility.

Assumption 1 (Outcome invariance for replay)**.** For the accounting replay, the realized payoff y_t attached to a stock-day is taken as fixed under the logged proposal and under the deployed hold action. Equivalently, the action does not change the subsequent price path in this testbed. This assumption is appropriate for the no-market-impact diagnostic abstraction, but it is not a claim about deployable trading or interactive environments.

Proposition 3 (Fixed-outcome replay accounting identity)**.** Under Assumption 1, for any fixed set of pre-fallback actions $\{\bar{a}_t\}_{t=1}^n$, the realized payoff difference between the deployed action and executing all logged proposals is

$$\Delta(\theta_D, \theta_w) = \sum_{t=1}^{n} [U(a_t^*, y_t) - U(\bar{a}_t, y_t)] = - \sum_{t:F_t(\theta_D,\theta_w)} \bar{a}_t \, y_t. \tag{7}$$

Therefore fallback improves realized payoff exactly when the rejected pre-fallback actions have negative aggregate signed payoff.

Proposition 3 is an accounting identity for a fixed outcome table. It does not provide a distribution-free selective-risk bound: such a bound would require an estimated score–risk relationship and would need to address the degenerate always-hold policy. For settings in which actions affect outcomes, randomized logging, a simulator, a structural model, or an off-policy estimator is required (Jiang & Li, 2016). We prove only the controller identities and the replay identity in Appendix A.

4 Experiments

4.1 Diagnostic protocol and evaluation environment

The primary evaluation uses the canonical MAG7 universe over a rolling January 2024–October 2025 horizon and contains 3,214 strictly matched stock-day decisions per variant. The replay records each pre-fallback action, forecast-model variable, trigger, deployed action, and realized payoff. All semantic inferences use greedy decoding (temperature 0.0) with fixed operational seeds. Textual context combines financial news and pricing records (Dong et al., 2024), filtered by ticker and timestamp before each decision. The retained 140-row file is a local audit slice; its field-level completeness counts are reported separately and are not used to replace or recompute the canonical results.

We compare five end-to-end variants and proposal-fixed controller replays. A1 full system enables the complete instrumented agent stack. A2 removes confidence weighting in its full pipeline and therefore is an end-to-end variant with its own proposal stream; its results are not controller-only estimates. A3 removes the multi-agent critique phase. A4 disables the verbal self-correction loop. A5 is the ungated operational agent without the forecast-guided fallback controller. A1-GR is a separate replay that holds the A1 proposal stream and logged outcomes fixed while replacing the adjusted divergence with the raw normalized divergence. Only A1 versus A1-GR and the A1 pre/post comparison are proposal-fixed controller contrasts. The retained 140-row slice additionally supports local coverage and data-completeness audits, which are labeled separately from the canonical results.

Cumulative signed utility is used only as a diagnostic tracing signal, not as a claim of deployable portfolio superiority. We report proposal coverage, deployed coverage, fallback composition, conditional directional error, simple matched-coverage controls, and the logged value of rejected actions. The 95% intervals use 2,000 paired two-sided percentile bootstrap replicates with ticker-month blocks; matched contrasts resample the same blocks for both quantities. The six-by-six threshold sweeps are sensitivity diagnostics, not validation-based threshold-selection evidence.

4.2 Replay diagnostic findings

Table 1 reports the canonical variant results and paired fixed-outcome contrasts; the whole-system and proposal-fixed estimands are interpreted separately below.

Table 1: Canonical matched outcomes and paired replay contrasts on 3,214 decisions. Utilities and contrasts are cumulative signed percentage points; intervals use paired ticker-month block bootstrap. A2 is end-to-end; A1-GR is a proposal-fixed derived replay, not an independent run.

Method	Matched decisions	Utility (pp)	Proposal utility (pp)	Fallback rate (%)
A1 full system	3,214	+6.37	+8.18	38.6
A2 no conf. weighting	3,214	+11.62	+36.63	51.8
A3 no debate	3,214	-3.20	+49.09	40.1
A4 no reflection	3,214	+17.16	+47.42	37.8
A5 ungated baseline	3,214	+23.56	+23.56	0.0
A1-GR fixed replay	3,214	+2.85	+8.18	50.6

Paired contrasts and 95% block-bootstrap intervals

Estimand	Contrast (pp)	95% CI	Interpretation
End-to-end A5 minus A1	+17.19	[4.80,30.90]	Combines proposal generation, participation, and controller changes.
A1 default gate minus no gate	-1.81	[-6.40,2.15]	Proposal-fixed controller contrast; inconclusive.
A1 low-confidence trigger	+3.70	[0.44,7.12]	Trigger-specific logged rejected-action value.
A1 divergence-only trigger	-5.48	[-9.07,-2.01]	Trigger-specific logged rejected-action value.
A1-GR raw replay minus no gate	-3.52	[-8.90,0.75]	Proposal-fixed raw-divergence replay.

The end-to-end gap combines proposal generation, participation, and controller changes; the proposal-fixed interval crosses zero, so this corpus does not resolve the default gate's average accounting effect. All contrasts are fixed-outcome diagnostics, not deployment-time causal effects.

In the frozen 1 August–31 October 2025 split ($n = 448$), the A1 deployed-minus-proposal contrast is -0.36 points [-1.41, 0.62], while the A5–A1 deployed-return gap is +2.79 points [0.61, 5.18] (Appendix Table 7). Thresholds were not retuned for this split. This temporal check does not rule out pretraining-data overlap with the evaluation period.

The difference between deployed and pre-fallback utility is the logged rejected-action value under Assumption 1. Trigger composition varies across variants (Appendix Table 6); for A1, the paired trigger-specific values are +3.70 points for low-confidence interventions and -5.48 for divergence-only interventions (Table 1). For A2, proposal utility is +36.63 points and deployed utility is +11.62, a -25.01-point within-variant contrast; this is an end-to-end profile, not a controller-only estimate.

Table 2: Simple matched controls on the canonical corpus. Coverage is the fraction of decisions with a nonzero deployed action; all values are descriptive fixed-outcome replay quantities.

Configuration	Utility (pp)	Coverage (%)	Scope
Always hold	0.00	0.00	Canonical non-trading null
A1 no gate	+8.18	79.40	Canonical A1 proposal stream
A1 low-confidence-only	+11.88	49.84	Reported proposal-fixed replay
A1 full gate	+6.37	46.92	$(\theta_D, \theta_w) = (0.30, 0.22)$
A1-GR raw-divergence replay	+2.85	37.18	Canonical proposal-fixed replay
A5 ungated	+23.56	80.68	Canonical end-to-end reference

The low-confidence-only replay exceeds the full-gate utility in this log, whereas the raw-divergence replay is lower. This pattern motivates separating gate components in the diagnostic; it is not evidence that a simplified rule is generally preferable.

4.3 COVERAGE AND CONDITIONAL RISK

Figure 2 plots selected operating points for the A1 controller. Across these configurations, deployed coverage ranges from 31.08% to 69.79%, while conditional directional risk ranges from 37.96% to 45.28% (Appendix Table 8). At the default thresholds, coverage is 46.92% and conditional risk is 41.91%. These descriptive points show the exposure–risk trade-off in this replay, not a calibrated selective-risk guarantee.

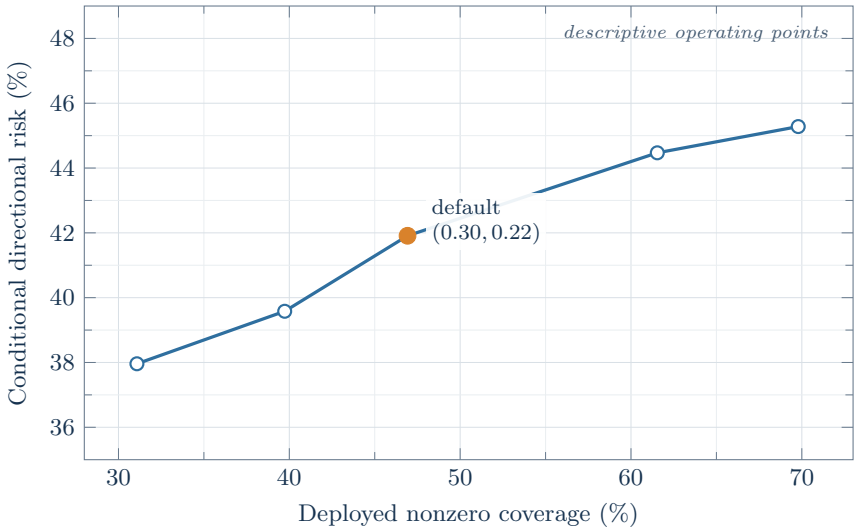

Figure 2: Selected risk–coverage operating points for the A1 controller on the canonical 3,214-row log. The default point $(\theta_D, \theta_w) = (0.30, 0.22)$ is labeled explicitly. Values are descriptive replay quantities and do not imply a selective-risk guarantee.

Risk-adjusted summaries provide a complementary descriptive view: at matched coverage, A5 has 46.15% conditional risk versus 41.91% for A1, while its annualized Sharpe is 1.04 versus 0.94 for A1 (Appendix Table 8). Since matching is descriptive rather than randomized and does not align proposal streams, these quantities do not identify an effect of the gate. At the trigger level, low-confidence and divergence-only interventions have opposite-signed rejected-action values (Table 1); aggregate utility alone therefore hides meaningful trigger heterogeneity.

For deployed nonzero actions, A1 has 632 directional errors out of 1,508 (41.91%, interval [39.43,44.47]), whereas A5 has 1,198 out of 2,593 (46.20%, interval [44.17,48.29]). The descriptive difference is -4.29 percentage points (interval [-7.92,-0.71]); because A1 and A5 have different proposal streams and coverage, it is not a controller-only error estimate. A1-GR retains the A1 proposal stream but deploys 1,195 nonzero actions, making it the relevant controller-only coverage comparison.

On the same A1 proposals, the raw-divergence A1-GR replay has 37.18% deployed coverage and +2.85 utility, compared with 46.92% and +6.37 for the default gate; its paired contrast against no gate is -3.52 points (95% CI [-8.90, 0.75]). This comparison changes participation as well as logged utility, and its interval includes zero; it does not establish that either divergence rule is preferable.

4.4 COVERAGE ACCOUNTING AND THRESHOLD SENSITIVITY

Table 3 makes the action-coverage denominators explicit. Natural holds account for 20.60% of A1 decisions before the controller; fallback flags are 38.61% of all decisions, and 1,044 nonzero proposals are overridden. Figure 3 shows the 15 recorded utility configurations for variants A1–A5. It is a descriptive summary, not the complete six-by-six sweep or evidence of validation-based threshold selection. The complete sweep boundaries and retained-slice coverage heatmaps are in Appendix D.

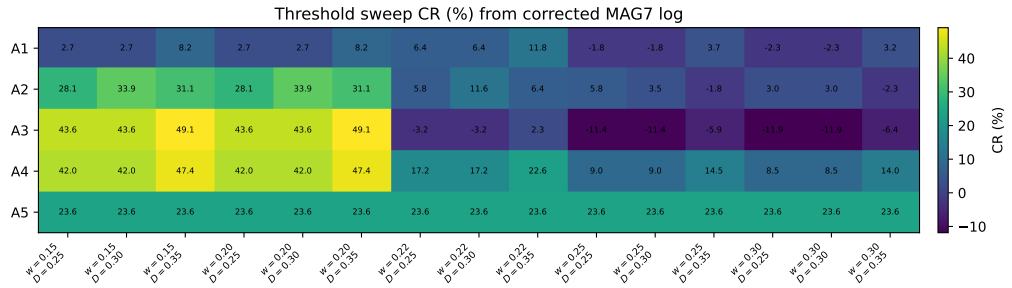

Figure 3: Threshold-sensitivity summary from the corrected MAG7 log. Rows are variants A1–A5 and columns are the 15 threshold configurations shown in the source plot; cell annotations report cumulative-return values (CR, %) as plotted. This is a descriptive summary, not a complete six-by-six grid or evidence of threshold optimality.

Table 3: Coverage decomposition on the canonical 3,214-row corpus. Natural holds are zero proposals before the controller. Fallback flags use all decisions as the denominator; nonzero overrides use nonzero proposals as the denominator; deployed coverage uses all decisions.

Variant	Natural hold	Proposal coverage	Fallback flags	Nonzero overrides	Deployed nonzero coverage
A1	662 (20.60%)	2,552 (79.40%)	1,241 (38.61%)	1,044 (40.91% of proposals)	1,508 / 3,214 (46.92%)
A1-GR fixed replay	662 (20.60%)	2,552 (79.40%)	50.6%	1,357 (53.17% of proposals)	1,195 / 3,214 (37.18%)
A5	621 (19.32%)	2,593 (80.68%)	0.0%	0	2,593 / 3,214 (80.68%)

4.5 Robustness and forecast-signal checks

The full six-by-six sweep is variant-dependent: utility ranges from 14.14 points for A1 to 60.95 for A3, while A5 is invariant at +23.56 across thresholds (Appendix Table 15). These are sensitivity bounds, not evidence for an optimal threshold. Removing decisions with operationally missing forecast features leaves A1 at +6.37 and A5 at +16.01; excluding the primary volatile entity yields +14.99 and +28.23, respectively (Appendix Table 17). The ordering persists in these exclusions, but the corpus remains limited to the same seven equities.

Within volatility terciles, A1 utility rises from +1.42 to +2.87 while A5 rises from +5.18 to +10.47; A1's fallback rate increases from 31.7% in the low-volatility bucket to 47.6% in the high-volatility bucket (Appendix Table 18). Bucket values are normalized within subsets and are not additive components of the aggregate. Across four chronological blocks, the A1–A5 utility gap changes sign (-15.35, +0.09, +5.62, and -7.55 points), showing that the aggregate ordering is not uniform over time (Appendix Table 21).

Forecast checks likewise support a bounded interpretation of the confidence signal. Chronos has CRPS 0.0128, 78.60% nominal 80% interval coverage, and confidence ECE 0.046; conformalization raises interval coverage to 80.71%, but the paired A1 gate contrast remains uncertain (-1.64 points, 95% CI [-6.19, 2.31]). TimesFM's rank association is 0.13 ($p = 0.018$), while its paired gate contrast also spans zero (-1.52 points, 95% CI [-5.93, 2.70]) (Appendix Table 11). These checks characterize logged signal quality; they do not validate deployment-time risk control.

5 DISCUSSION

5.1 INTERPRETING LOGGED REJECTED-ACTION VALUE

The A5–A1 end-to-end gap combines proposal generation and participation, while the A1 proposal-fixed replay reports a logged accounting contrast under fixed outcomes. Their different magnitudes and the opposite-signed trigger values motivate Fixed-Outcome Replay as an accounting diagnostic, not a causal ranking of gate designs. A2–A4 change upstream components and therefore describe whole-system profiles.

5.2 SCOPE AND LIMITATIONS

Evidence covers 3,214 MAG7 stock-days from January 2024 to October 2025; it does not establish robustness across sectors or regimes. Outcomes are fixed across candidate actions, and trading costs, slippage, market impact, and position sizing are omitted, so the contrasts are accounting diagnostics rather than deployment effects. Bootstrap intervals quantify uncertainty within this corpus. Confidence is operational, and Chronos checks do not establish calibration under drift or forecaster invariance. Action-dependent effects require randomized data, a validated simulator, a structural model, or an appropriate off-policy estimator.

The auxiliary bullish-window stress test also cautions against an alpha interpretation: passive always-buy reaches +91.87%, compared with +24.47% for A5 on that separate sequence (Appendix Table 22). Because the window is deliberately bullish and separate from the canonical matched corpus, this is a contextual benchmark rather than evidence of broad baseline dominance.

For an applied audit, proposal coverage, fallback rate, deployed coverage, conditional risk, and trigger-level rejected-action value should be read together. They answer different questions: a lower error rate among accepted actions can coincide with substantially lower participation, while the signed value of rejected proposals can vary by trigger. Collapsing these quantities into one aggregate safety score would obscure the very distinctions FOR is designed to make. In this study, the whole-system A5–A1 gap and the A1 proposal-fixed replay are therefore reported side by side, with the latter interpreted only under the fixed-outcome assumption.

6 CONCLUSION

Fixed-Outcome Replay makes logged gate behavior auditable under fixed outcomes. In the 3,214-decision MAG7 corpus, trigger values differ by type and the default proposal-fixed contrast is inconclusive. Deployment claims require action-dependent evidence.

USE OF LLMS

OpenAI Codex was used to provide feedback on the manuscript's framing and methodology, discuss interpretations of results supplied by the authors, revise manuscript text, and assist with figure and document preparation. The authors made the final decisions about the study and are responsible for verifying and approving the data, analyses, references, and claims in the submission.

REFERENCES

Sami Achour, Yassine Bouher, Duong Nguyen, and Nicolas Chesneau. Foundation models for time series forecasting: Application in conformal prediction. *arXiv preprint arXiv:2507.08858*, 2025. URL https://arxiv.org/abs/2507.08858.

Yihao Ang, Yifan Bao, Lei Jiang, Jiajie Tao, Anthony K. H. Tung, Lukasz Szpruch, and Hao Ni. Structured agentic workflows for financial time-series modeling with llms and reflective feedback. *arXiv preprint arXiv:2508.13915*, 2025. URL https://arxiv.org/abs/2508.13915.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=33XGfHLtZg.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2403.07815.

Somnath Banerjee, Sayan Layek, Soham Tripathy, Shanu Kumar, Animesh Mukherjee, and Rima Hazra. Safeinfer: Context adaptive decoding time safety alignment for large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 27188–27196, 2025. doi: 10.1609/aaai.v39i26.34927. URL https://ojs.aaai.org/index.php/AAAI/article/view/34927.

Amrita Bhattacharjee, Shaona Ghosh, Traian Rebedea, and Christopher Parisien. Towards inference-time category-wise safety steering for large language models. *arXiv preprint arXiv:2410.01174*, 2024. URL https://arxiv.org/abs/2410.01174.

Jonathan Brogaard, Terrence Hendershott, and Ryan Riordan. High-frequency trading and price discovery. *The Review of Financial Studies*, 27(8):2267–2306, 2014. doi: 10.1093/rfs/hhu032.

Baiting Chen, Zhimei Ren, and Lu Cheng. Conformalized time series with semantic features. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3859. URL https://papers.neurips.cc/paper_files/paper/2024/hash/dbfb7b1443583fc7ab87e8b1b4f48c9c-Abstract-Conference.html.

Yanxu Chen, Zijun Yao, Yantao Liu, Amy Xin, Jin Ye, Jianing Yu, Lei Hou, and Juanzi Li. StockBench: Can LLM agents trade stocks profitably in real-world markets? *arXiv preprint arXiv:2510.02209*, 2025. URL https://arxiv.org/abs/2510.02209.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 10148–10167. PMLR, 2024. URL https://proceedings.mlr.press/v235/das24c.html.

Zihan Dong, Xinyu Fan, and Zhiyuan Peng. Fnspid: A comprehensive financial news dataset in time series. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4918–4927, 2024. URL https://arxiv.org/abs/2402.06698.

Elliot L. Epstein, John Winnicki, Thanawat Sornwanee, and Rajat Dwaraknath. Llms are overconfident: Evaluating confidence interval calibration with fermieval. *arXiv preprint arXiv:2510.26995*, 2025. URL https://arxiv.org/abs/2510.26995.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. Don't hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 14664–14690, 2024. doi: 10.18653/v1/2024.acl-long.786.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 2151–2159, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Tobias Groot and Matias Valdenegro-Toro. Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing*, pp. 145–171, 2024. doi: 10.18653/v1/2024.trustnlp-1.13.

Rima Hazra, Sayan Layek, Somnath Banerjee, and Soujanya Poria. Safety arithmetic: A framework for test-time safety alignment of language models by steering parameters and activations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 21759–21776. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.1212. URL https://aclanthology.org/2024.emnlp-main.1212/.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, pp. 652–661, 2016. URL https://proceedings.mlr.press/v48/jiang16.html.

Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J. Bell. Abstentionbench: Reasoning llms fail on unanswerable questions. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5729. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/fb122bfc3f0127a94ded048b5b03496f-Abstract-Datasets_and_Benchmarks_Track.html. Datasets and Benchmarks Track.

Abhishek Kumar, Robert Morabito, Sanzhar Umbet, Jad Kabbara, and Ali Emami. Confidence under the hood: An investigation into the confidence-probability alignment in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 315–334. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.20. URL https://aclanthology.org/2024.acl-long.20/.

Chengzu Li, Han Zhou, Goran Glavaš, Anna Korhonen, and Ivan Vulić. Large language models are miscalibrated in-context learners. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11575–11596, 2025a. doi: 10.18653/v1/2025.findings-acl.603.

Feifei Li, Haote Xu, Haodi Xu, Yinhao Liu, and Xinghao Ding. Cvc: Further aligning llms via cross-view correction for time series forecasting. *Knowledge-Based Systems*, 326:113957, 2025b. doi: 10.1016/j.knosys.2025.113957. URL https://www.sciencedirect.com/science/article/pii/S0950705125010020.

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

Xu Liu, Juncheng Liu, Gerald Woo, Taha Aksu, Yuxuan Liang, Roger Zimmermann, Chenghao Liu, Junnan Li, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. Moirai-moe: Empowering time series foundation models with sparse mixture of experts. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 38940–38962. PMLR, 2025b. URL https://proceedings.mlr.press/v267/liu25an.html.

Nishanth Madhusudhan, Sathwik Tejaswi Madhusudhan, Vikas Yadav, and Masoud Hashemi. Do llms know when to not answer? investigating abstention abilities of large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 9329–9345. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.coling-main.627/.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 7076–7087, 2020. URL https://proceedings.mlr.press/v119/mozannar20b.html.

Vy Nguyen, Ziqi Xu, Jeffrey Chan, Estrid He, Feng Xia, and Xiuzhen Zhang. Hallucinate less by thinking more: Aspect-based causal abstention for large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 32555–32563, 2026. doi: 10.1609/aaai.v40i38.40532. URL https://ojs.aaai.org/index.php/AAAI/article/view/40532.

Edward Phillips, Fredrik K. Gustafsson, Sean Wu, Anshul Thakur, and David A. Clifton. Entropy alone is insufficient for safe selective prediction in llms. *arXiv preprint arXiv:2603.21172*, 2026. URL https://arxiv.org/abs/2603.21172.

Eghbal Rahimikia, Hao Ni, and Weiguan Wang. Re(visiting) time series foundation models in finance. *arXiv preprint arXiv:2511.18578*, 2025. URL https://arxiv.org/abs/2511.18578.

Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Hena Ghonia, Rishika Bhagwatkar, Arian Khorasani, Mohammad Javad Darvishi Bayazi, George Adamopoulos, Roland Riachi, Nadhir Hassen, Marin Biloš, Sahil Garg, Anderson Schneider, Nicolas Chapados, Alexandre Drouin, Valentina Zantedeschi, Yuriy Nevmyvaka, and Irina Rish. Lag-llama: Towards foundation models for probabilistic time series forecasting. *arXiv preprint arXiv:2310.08278*, 2024. URL https://arxiv.org/abs/2310.08278.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2303.11366.

Dennis Ulmer, Martin Gubri, Hwaran Lee, Sangdoo Yun, and Seong Oh. Calibrating large language models using their generations only. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15440–15459. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.824. URL https://aclanthology.org/2024.acl-long.824/.

Pengyu Wang, Dong Zhang, Linyang Li, Chenkun Tan, Xinghao Wang, Ke Ren, Botian Jiang, and Xipeng Qiu. InferAligner: Inference-time alignment for harmlessness through cross-model guidance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10533–10549, 2024. doi: 10.18653/v1/2024.emnlp-main.585. URL https://arxiv.org/abs/2401.11206.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. Know your limits: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics*, 13:529–556, 2025. doi: 10.1162/tacl_a_00754.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 53140–53164. PMLR, 2024. URL https://proceedings.mlr.press/v235/woo24a.html.

Jiancong Xiao, Bojian Hou, Zhanliang Wang, Ruochen Jin, Qi Long, Weijie J. Su, and Li Shen. Restoring calibration for aligned large language models: A calibration-aware fine-tuning approach. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 68364–68390. PMLR, 2025. URL https://proceedings.mlr.press/v267/xiao25b.html.

Yijia Xiao, Edward Sun, Di Luo, and Wei Wang. Tradingagents: Multi-agents llm financial trading framework. *arXiv preprint arXiv:2412.20138*, 2024. URL https://arxiv.org/abs/2412.20138.

Zeguan Xiao, Diyang Dou, Boya Xiong, Yun Chen, and Guanhua Chen. Enhancing uncertainty estimation in llms with expectation of aggregated internal belief. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34043–34051, 2026. doi: 10.1609/aaai.v40i40.40698. URL https://ojs.aaai.org/index.php/AAAI/article/view/40698.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. Fingpt: Open-source financial large language models. *arXiv preprint arXiv:2306.06031*, 2023. URL https://arxiv.org/abs/2306.06031.

Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yuechen Jiang, Yupeng Cao, Zhi Chen, Jordan W. Suchow, Zhenyu Cui, Rong Liu, Zhaozhuo Xu, Denghui Zhang, Koduvayur Subbalakshmi, Guojun Xiong, Yueru He, Jimin Huang, Dong Li, and Qianqian Xie. Fincon: A synthesized llm multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/f7ae4fe91d96f50abc2211f09b6a7e49-Abstract-Conference.html.

Qingcheng Zeng, Weihao Xuan, Leyang Cui, and Rob Voigt. Thinking out loud: Do reasoning models know when they're right? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1394–1407. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.73. URL https://aclanthology.org/2025.emnlp-main.73/.

Jingyu Zhang, Ahmed Elgohary Ghoneim, Ahmed Magooda, Daniel Khashabi, and Benjamin Van Durme. Controllable safety alignment: Inference-time adaptation to diverse safety requirements. In *International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=ERce2rgMQC.

Mozhi Zhang, Mianqiu Huang, Rundong Shi, Linsen Guo, Chong Peng, Peng Yan, Yaqian Zhou, and Xipeng Qiu. Calibrating the confidence of large language models by eliciting fidelity. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2959–2979. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.173. URL https://aclanthology.org/2024.emnlp-main.173/.

Xiaohan Zhang, Tian Gao, Mingyue Cheng, Bokai Pan, Ze Guo, Yaguo Liu, Xiaoyu Tao, and Qi Liu. Alphacast: A human wisdom-llm intelligence co-reasoning framework for interactive time series forecasting. *arXiv preprint arXiv:2511.08947*, 2025b. URL https://arxiv.org/abs/2511.08947.