Simulated Ignorance Fails: A Systematic Study of LLM Behaviors on Forecasting Problems Before Model Knowledge Cutoff

Zehan Li[1] , **Yuxuan Wang**[2] , **Ali El Lahib**[2] , **Ying-Jieh Xia**[2] , **Xinyu Pi**[2]
[1]University of Chicago
[2]University of California, San Diego
zehan@uchicago.edu, {waw009, aellahib, yix050, xpi}@ucsd.edu

Abstract

Evaluating LLM forecasting capabilities is constrained by a fundamental tension: prospective evaluation offers methodological rigor but prohibitive latency, while retrospective forecasting (RF)—evaluating on already-resolved events—faces rapidly shrinking clean evaluation data as SOTA models possess increasingly recent knowledge cutoffs. **Simulated Ignorance (SI)**, prompting models to suppress pre-cutoff knowledge, has emerged as a potential solution. We provide the first systematic test of whether SI can approximate **True Ignorance (TI)**. Across 477 competition-level questions and 9 models, we find that SI fails systematically: (1) cutoff instructions leave a 52% performance gap between SI and TI; (2) chain-of-thought reasoning fails to suppress prior knowledge, even when reasoning traces contain no explicit post-cutoff references; (3) reasoning-optimized models exhibit *worse* SI fidelity despite superior reasoning trace quality. These findings demonstrate that prompts cannot reliably "rewind" model knowledge. We conclude that RF on pre-cutoff events is methodologically flawed; we recommend against using SI-based retrospective setups to benchmark forecasting capabilities.

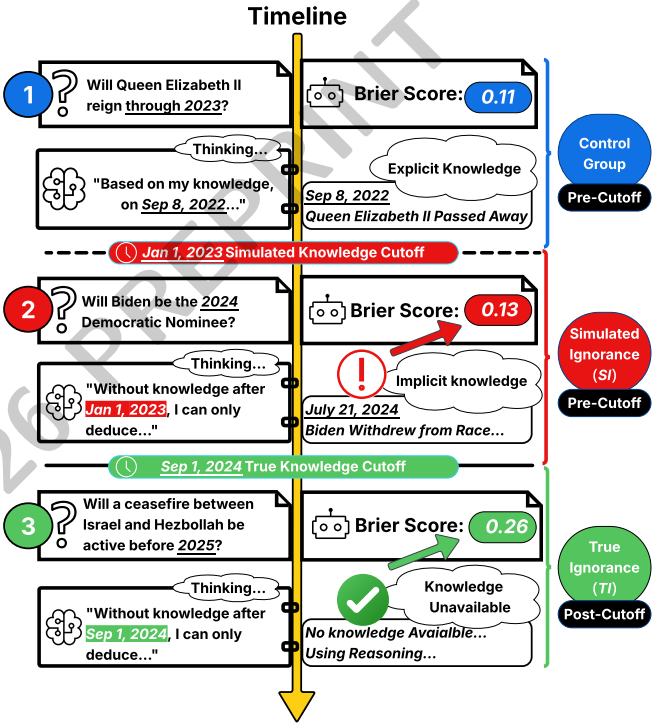

Figure 1: **Simulated Ignorance (SI) fails to suppress Implicit Knowledge. (1) Control Group:** Models accurately predict past events using available **Explicit Knowledge** (Brier: 0.11). **(2) Simulated Ignorance:** Despite a prompted cutoff (Jan 2023), the model retains **Implicit Knowledge** of future outcomes (e.g., Biden's withdrawal). As indicated by the red arrow, this leakage improves the forecast (Brier: 0.13) **(3) True Ignorance (TI):** When knowledge is genuinely **Unavailable** (e.g., post-training ceasefire), the model must rely on pure reasoning, resulting in significantly higher uncertainty (0.26). The SI–TI gap confirms that prompts cannot reliably "rewind" the model's knowledge state.

1 Introduction

LLMs are increasingly deployed as forecasting tools, with applications spanning geopolitical prediction, financial analysis, and scientific discovery [Zou *et al.*, 2022; Halawi *et al.*, 2024; Pratt *et al.*, 2024; Schoenegger *et al.*, 2024; Phan *et al.*, 2024]. The appeal is clear: models trained on vast corpora of human knowledge might internalize patterns that enable them to anticipate future developments. However, evaluating such forecasting capabilities presents a fundamental methodological challenge.

Prospective evaluation—deploying models to make predictions about genuinely future events and waiting for outcomes—offers the gold standard for assessing forecasting ability, but suffers from prohibitive latency. Years may pass before predictions can be validated, making rapid model evaluation impractical. Consequently, the field has largely converged on *retrospective forecasting (RF)* as a compromise: evaluating models on historical events that have already resolved [Zou *et al.*, 2022; Phan *et al.*, 2024]. When such events postdate the model's training cutoff, the model operates under *True Ignorance (TI)*—it genuinely lacks knowledge of the outcome [Schoenegger *et al.*, 2024; Halawi *et al.*, 2024].

However, TI-based evaluation is becoming increasingly

difficult. As Paleka *et al.* [2025] highlight, state-of-the-art models now possess knowledge cutoffs extending into late 2024 or early 2025. This means most historical events have already been seen during training—leaving very few post-cutoff events available for evaluation. The most capable models are precisely those with the most recent cutoffs, and therefore the fewest usable evaluation questions. The problem is worse for model comparison: if we want to evaluate multiple models on the same questions, those questions must postdate *all* models' cutoffs, further limiting the pool.

One proposed solution is to extend RF to pre-cutoff events by instructing models to suppress their knowledge—a paradigm we term ***Simulated Ignorance (SI)*** [Pawelczyk *et al.*, 2024; Gao *et al.*, 2025]. **This is not a hypothetical practice but a recurring assumption in LLM forecasting evaluation and development**: Autocast [Zou *et al.*, 2022] prompts models with temporal cutoffs to test them on events predating training; Phan *et al.* [2024] report strong LLM forecasting on pre-cutoff data without SI controls; and Halawi *et al.* [2024]'s 'near-human" forecasting system uses date-filtered retrieval and tunes components on already-resolved questions. The risk is not confined to the final test set: even when evaluation is restricted to post-cutoff questions, prompts, retrieval filters, aggregation rules, and hyperparameters may be selected on pre-cutoff resolved data, implicitly assuming that models do not exploit known outcomes during optimization. If SI fails, such pre-cutoff tuning cannot serve as a clean proxy for genuine forecasting—even when the held-out evaluation is temporally cleaner. Moreover, every static benchmark becomes an implicit SI setup the moment newer models' cutoffs pass the questions' resolution dates—already happening with current SOTA. For instance, to evaluate forecasting on the 2024 U.S. presidential election, one might instruct the model: '*use only knowledge available before January 1, 2023*." If successful, this approach would unlock vast repositories of historical data for forecasting evaluation, circumventing this data scarcity.

But does ***SI*** actually work? The test is straightforward: if prompting successfully suppresses prior knowledge, then forecasting performance under ***SI*** should match performance under ***TI***. Conversely, if models perform significantly better on questions they "should not know," they are likely benefiting from leaked knowledge, invalidating ***SI***-based RF.

Our paper provides the first systematic empirical test of this assumption. Across 477 competition forecasting questions evaluated on 9 models—partitioned into pre-cutoff questions (evaluated under SI) and post-cutoff questions (evaluated under TI)—we demonstrate that ***SI*** fails systematically.

Our investigation proceeds through three increasingly strong interventions. **(1)** We evaluate instruction prompting with explicit temporal constraints; models exhibit a 52% **performance gap** between ***SI*** and ***TI***, with substantially inflated accuracy and confidence on questions whose answers they were instructed to suppress. **(2)** We elicit structured reasoning, hypothesizing that explicit reasoning might force models to "show their thinking" and adhere to temporal constraints. While structured reasoning narrows the gap, it fails to close it. More concerningly, models produce reasoning traces that appear fully compliant with the cutoff instruction, yet their forecasting performance still reflect leaked knowledge—suggesting that surface-level inspection of reasoning cannot reliably diagnose leakage. **(3)** We test reasoning-optimized models trained with reinforcement learning, hypothesizing they might exhibit stronger instruction-following in temporal suppression tasks [DeepSeek-AI, 2025]. Paradoxically, these models exhibit *worse* SI fidelity despite superior reasoning quality and overall forecasting performance.

All of our intervention attempts failed to rewind model knowledge. We conclude that **SI-based retrospective evaluation**—using prompts to simulate ignorance about pre-cutoff events—is **methodologically flawed**: ***SI*** cannot approximate ***TI***, and we recommend against benchmarking to pre-cutoff events since knowledge leakage cannot be eradicated.

2 Related Work

2.1 LLM Forecasting and Retrospective Evaluation

Earlier RF benchmarks adopted divergent strategies for handling the leakage risk: some evaluated on pre-cutoff events with date-restriction prompts [Zou *et al.*, 2022; Halawi *et al.*, 2024; Phan *et al.*, 2024], while others restricted to post-cutoff events to ensure genuine ignorance [Schoenegger *et al.*, 2024]. The pre-cutoff approach—***Simulated Ignorance (SI)***—is the *de facto* paradigm in LLM forecasting, since it preserves access to the much larger pool of already-resolved questions.

Recent benchmarks have quietly moved away from pre-cutoff evaluation, constructing prospective-only datasets without explicit justification [Karger *et al.*, 2025; Paleka *et al.*, 2025; FutureSearch *et al.*, 2025]. This implicit retreat signals growing unease, but no prior work has directly tested whether SI can approximate ***True Ignorance (TI)*** or quantified how badly it fails. We make this implicit concern explicit through systematic empirical comparison.

2.2 Temporal Contamination and Surface Compliance

The failure of SI to approximate TI reflects a broader phenomenon: models' forecasts may benefit from leaked knowledge even when instructed to ignore it. This connects to work on data leakage in NLP evaluation more generally [Sainz *et al.*, 2023; Balloccu *et al.*, 2024; Magar and Schwartz, 2022]. In domains such as finance, prior studies demonstrate that models benefit from leaked knowledge under retrospective evaluation [Lopez-Lira *et al.*, 2025]. A related line of work investigates prompt-based "unlearning": whether instructions can suppress explicit mentions of known facts [Pawelczyk *et al.*, 2024; Gao *et al.*, 2025].

Our setting exposes a critical distinction: *surface compliance* versus *implicit knowledge retention*. A model may obediently avoid mentioning post-cutoff facts in its reasoning while its forecasting performance still reflects leaked knowledge. We therefore measure leakage via forecasting scores (Brier scores) rather than reasoning content.

2.3 Reasoning Faithfulness and Post-Hoc Rationalization

If surface compliance does not guarantee knowledge suppression, a natural question arises: can we trust reasoning traces to reveal when models are "cheating"? Chain-of-thought (CoT) prompting is widely adopted to improve both interpretability and performance [Wei *et al.*, 2022; Kojima *et al.*, 2022]. However, a growing literature shows that reasoning traces can be misleading: models may produce convincing explanations that do not reflect their actual decision process [Turpin *et al.*, 2023; Lanham *et al.*, 2023; Paul *et al.*, 2024].

This concern is amplified for reasoning-optimized models (e.g., O1, R1), trained via reinforcement learning to produce extended traces [OpenAI, 2024; DeepSeek-AI, 2025]. [Chen *et al.*, 2025] demonstrate unfaithful reasoning in such models on mathematical tasks, where errors in CoT can be detected by verifying logical consistency. Forecasting offers no such verification: there is no "correct" reasoning path, so a cutoff-compliant trace may still mask leaked knowledge.

Our results provide forecasting-specific evidence for this concern. Reasoning-optimized models exhibit the largest SI–TI gaps despite producing the cleanest traces—suggesting RL training teaches models to rationalize predictions rather than derive them transparently.

3 Experimental Setup

3.1 Problem Formulation

We test whether ***Simulated Ignorance (SI)***—prompting models to use only knowledge before a specified date—can make retrospective forecasting on pre-cutoff events behave like genuine forecasting on post-cutoff events ***(True Ignorance, TI)***. Each model m has a reported knowledge cutoff $K(m)$. We prompt models with a simulated cutoff C preceding all question resolutions, and test whether this eliminates outcome-level dependence: if SI works, models should perform equally on questions resolved before versus after $K(m)$.

3.2 Data

We evaluate 9 models on 477 resolved binary questions from Metaculus tournament archives (2023–2025) [Metaculus, 2025].[1] Models span two categories: 6 *non-reasoning* models and 3 *reasoning-optimized* models trained with reinforcement learning. Questions are stratified across six domains (Business, Economics/Finance, Social/institutional, Entertainment, Technology/innovation, Politics/Geopolitics) to ensure balanced coverage.

To ensure reliability, we run each model-question pair 3 times per condition and average the predictions. Sample sizes vary slightly across models (Table 1) due to occasional API failures or malformed outputs; we exclude such cases from analysis. We macro-average all metrics over questions.

[1]Binary outcomes admit clean Brier-based SI–TI comparison; see the supplementary for alternative formulations.

Model	Cutoff	n (Pre/Post)
Reasoning (R)		
GPT-5.1	Sept '24	201 / 276
Gemini-2.5-Pro	Jan '25	201 / 276
DeepSeek-R1	Jan '25	201 / 276
Non-Reasoning (NR)		
Claude 3.5 Sonnet	Apr '24	201 / 274
Gemini-2.5-Flash	Jan '25	201 / 273
Llama 4 Maverick	Aug '24	201 / 274
DeepSeek-V3.2-Exp	Dec '24	201 / 273
OpenAI-OSS-120b	Jun '24	201 / 276
OpenAI-OSS-20b	Jun '24	199 / 276

Table 1: Model metadata. R = reasoning-optimized via RLVR; NR = standard instruction-tuned. Cutoffs from official documentation. Sample sizes vary slightly due to API failures.The Model knowledge cutoff date is available in [Wang, 2025]

3.3 Question Partition

To ensure all models are evaluated on the same questions, we use a *fixed* partition rather than per-model splits. Let $K_{\min}$ and $K_{\max}$ denote the earliest and latest knowledge cutoffs across all evaluated models. Each question q has resolution date $R(q)$, and we partition questions as follows:

- **Pre-cutoff**: $R(q) < K_{\min}$. The outcome predates *all* models' cutoffs; every model has likely seen the answer during training.
- **Post-cutoff**: $R(q) > K_{\max}$. The outcome postdates *all* models' cutoffs; no model has seen the answer.

When we add a cutoff instruction, Pre-cutoff questions become the **SI** condition (models must suppress knowledge they have), and Post-cutoff questions become the **TI** condition (models genuinely lack knowledge). The core test is whether performance under SI matches performance under TI.

This design ensures all models face the same question sets, but introduces a potential confound: models with later cutoffs are temporally closer to Post-cutoff question resolutions, and may benefit from more recent background context even without knowing answers. To verify that observed SI–TI gaps reflect genuine knowledge leakage rather than temporal proximity effects, we conduct robustness checks examining performance as a function of distance between cutoff and resolution (§5).

3.4 Prompting Conditions

We cross *reasoning elicitation* with *cutoff constraint* in a 2×2 design (Table 2): We fix $C = \texttt{2023-01-01}$, preceding all question resolutions.

	No Cutoff	**Cutoff at C**
Direct	G1	G1′
Reasoning (CoT)	G2	G2′

Table 2: Four prompting conditions formed by crossing reasoning elicitation (direct vs. CoT) with cutoff constraint (none vs. instructed cutoff C).

Under G1 and G2 (no cutoff instruction), we observe baseline performance on Pre-cutoff and Post-cutoff questions. Under G1′ and G2′ (with cutoff instruction), Pre-cutoff questions become the SI condition and Post-cutoff questions become the TI condition. This design allows us to isolate the effects of cutoff instructions (G1 vs. G1′) and reasoning elicitation (G1′ vs. G2′) on the SI–TI gap.

3.5 Metrics

Models output probability $p \in [0, 1]$; ground truth is $y \in \{0, 1\}$. We report Brier score $B(p, y) = (p - y)^2$ (lower is better; uninformed baseline $p = 0.5$ yields $B = 0.25$).

Under cutoff instruction conditions (G1′, G2′), we define the **SI–TI Gap**:

$$\Delta_g = \mathbb{E}_{q \in \text{Post}}[B_g(q)] - \mathbb{E}_{q \in \text{Pre}}[B_g(q)] \quad (1)$$

A large positive Δ_g indicates substantially better Pre-cutoff performance under SI—evidence that models fail to suppress leaked knowledge. If SI were effective, $\Delta_g \approx 0$.

4 Results

We test whether prompt-based interventions can make SI approximate TI. Across 9 models and three intervention layers, we find that all interventions reduce explicit leakage but fail to close the SI–TI gap.

4.1 Layer 1: Cutoff Instructions Reduce but Do Not Eliminate the Gap

Fig. 2A shows a consistent pattern across all 9 models: adding a cutoff instruction (G1′) reduces knowledge leakage on Pre-cutoff questions, moving it toward the Post-cutoff baseline—but a substantial gap remains. Quantitatively, the cutoff closes 48% of the SI–TI gap, leaving 52% unexplained (Fig. 2B). All models still show significant gaps under the cutoff prompt. SI therefore remains an unreliable proxy for TI even with explicit temporal constraints.

The SI–TI gap varies substantially across domains: cutoff instructions close 91% of the gap in the most responsive domain (BUS) but only 33% in the least responsive (POL-GEO). This pattern suggests leakage severity correlates with event salience—geopolitical events like elections receive extensive coverage and become deeply encoded, making them harder to suppress than lower-salience business metrics (see supplementary material for detailed domain-level analysis).

4.2 Layer 2: CoT Narrows the Gap but Clean Traces Hide Leakage

We next test whether chain-of-thought (CoT) prompting helps models adhere to temporal constraints by forcing them to "show their work."

CoT narrows the gap asymmetrically. Without a cutoff instruction (Fig. 3A), CoT improves Post-cutoff performance but *worsens* Pre-cutoff performance, narrowing the baseline gap from both directions. This asymmetry suggests that under direct prompting, models exploit a retrieval shortcut on Pre-cutoff questions—directly accessing memorized outcomes without explicit reasoning. CoT disrupts this shortcut by requiring verbalization, degrading Pre-cutoff performance while benefiting Post-cutoff where no shortcut exists.

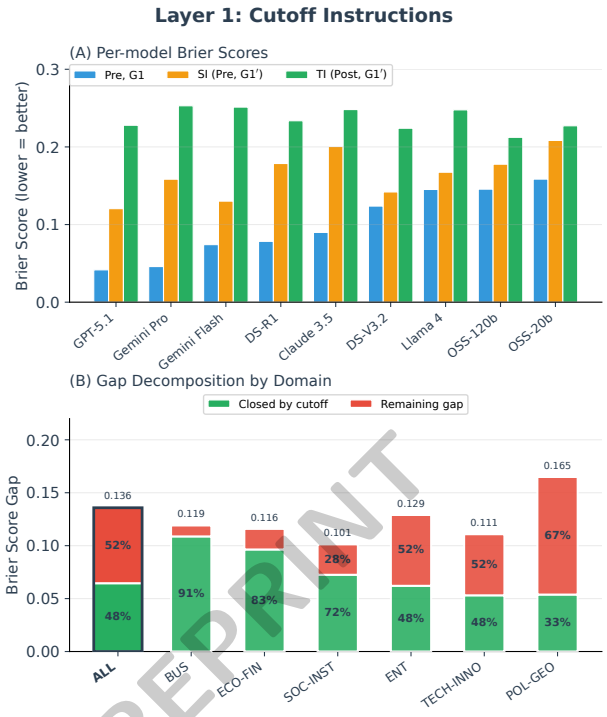

Figure 2: **Layer 1: Cutoff instructions reduce but do not eliminate the SI–TI gap. (A)** Brier scores across 9 models under three conditions: Pre-cutoff baseline (Pre, G1), SI (Pre, G1′), and TI (Post, G1′). **(B)** Gap decomposition by domain: portion closed by cutoff (green) vs. residual SI–TI gap (red). Lower Brier = better.

With a cutoff instruction (Fig. 3B), switching from direct to CoT prompting (G1′ → G2′) produces only small changes on both SI and TI. The cutoff instruction appears to already disrupt the retrieval shortcut, leaving less room for CoT to further reduce leakage.

Clean traces do not imply no leakage. We audit CoT traces on Pre-cutoff questions for logical consistency and explicit post-cutoff references.[2] Trace quality improves from G2 to G2′ (Fig. 4B): cutoff prompts make traces *look* cleaner. However, cleaner traces coexist with residual SI–TI gaps—the absence of explicit temporal violations does not guarantee the absence of leaked knowledge in the forecast.

4.3 Layer 3: Reasoning-Optimized Models Show Cleaner Traces but Larger Gaps

Reasoning-optimized models (e.g., O1, R1) are trained via reinforcement learning to produce structured reasoning and follow instructions consistently. We test whether this training improves SI fidelity.

[2]Logic consistency checks for internal contradictions, mathematical errors, and reasoning-conclusion alignment. Cutoff compliance detects references to information only knowable after the specified cutoff. See supplementary material for more details.

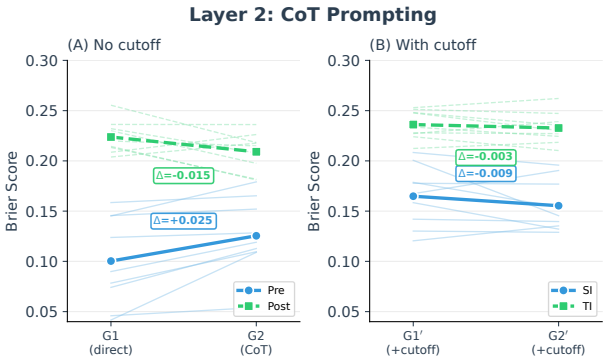

Figure 3: **Layer 2: CoT prompting effects.** (A) Without cutoff instruction: CoT worsens Pre-cutoff but improves Post-cutoff, narrowing the baseline gap. (B) With cutoff instruction: G1′ → G2′ yields small changes on both SI and TI.

Group	Logic (%)	Clean (%)
Reasoning-optimized	99.4	98.8
Non-reasoning	96.7	98.2

Table 3: **Layer 3: Trace audit under G2′.** Reasoning-optimized models produce more coherent and cutoff-clean traces under G2′.

Better TI performance, worse SI fidelity. Fig. 5A shows that reasoning-optimized models achieve strong TI performance across all prompting conditions. Their traces also appear more compliant: under G2′, they show higher logic-consistency (99.4% vs. 96.7%) and cutoff-compliance (98.8% vs. 98.2%) than non-reasoning models (Table 3).

However, reasoning-optimized models exhibit *larger* SI–TI gaps than non-reasoning models (Fig. 5B): 0.086 vs. 0.064 under G1′, and 0.097 vs. 0.067 under G2′. This dissociates forecasting capability from leakage resistance: models can forecast well under TI while failing to suppress prior knowledge under SI.

This pattern is consistent with RL training creating tension between two objectives: producing compliant traces and arriving at correct answers. Under SI, models learn to construct reasoning that avoids explicit violations while still guiding toward the encoded answer—appearing compliant without being compliant (see supplementary material for extended analysis).

4.4 Summary

Across all three interventions, we observe a common failure mode: interventions operate at the level of *model outputs*—what the model says—rather than *model knowledge*—what the model knows.

Cutoff instructions suppress explicit temporal references, CoT prompting produces coherent reasoning traces, and reasoning-optimized training yields high trace compliance—yet all three leave residual SI–TI gaps. A model can avoid every detectable form of leakage while implicit knowledge still influences its forecasting performances.

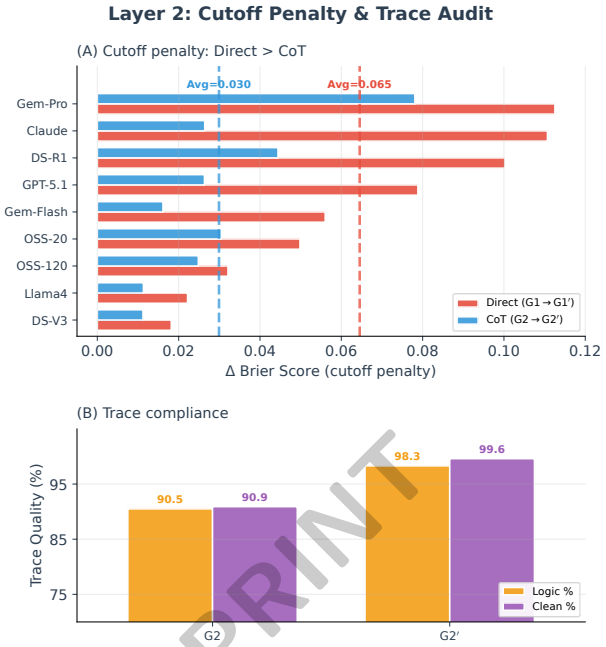

Figure 4: **Layer 2: Trace compliance.** (A) Cutoff penalty on Pre-cutoff is larger for direct prompting than CoT. (B) Trace audit metrics improve from G2 to G2′.

5 Robustness & Qualitative Analysis

Our main finding—that SI fails to approximate TI—could be challenged by an alternative explanation: perhaps post-cutoff questions are simply *harder* to forecast, independent from data leakage. We address this concern with experiments.

5.1 Difficulty Control: The Gap Is Model-Specific, Not Calendar-Specific

If difficulty (not leakage) causes the SI–TI gap, all models should struggle on the same questions at the same calendar time. If leakage causes the gap, performance should degrade at each model's *own* knowledge cutoff, regardless of calendar time. We test this with three complementary analyses.

Cross-model natural experiment. We exploit the fact that different models have different cutoffs. Q4 2024 questions provide a decisive test: they are **Pre-cutoff** for DeepSeek-V3.2 (cutoff: Jan 2025) but **Post-cutoff** for GPT-5.1 (cutoff: Sept 2024). If these questions were inherently harder, *both* models should struggle.

Model	**Q4'24 Status**	**Brier**	**Horizon**
GPT-5.1 (Sept'24)	TI	.204	1–3 mo
DeepSeek-V3.2 (Jan'25)	SI	.136	1–3 mo

Table 4: **Cross-model natural experiment.** Same questions, same horizon, different SI/TI status. The gap (+0.068) tracks knowledge cutoff, not question difficulty.

The same questions yield Brier 0.136 for DeepSeek but 0.204 for GPT-5.1. The only variable is whether the questions

Figure 5: **Layer 3: Reasoning-optimized models show cleaner traces but larger SI–TI gaps.** (A) Brier scores across prompting conditions. (B) SI–TI gap comparison: reasoning-optimized vs. non-reasoning models.

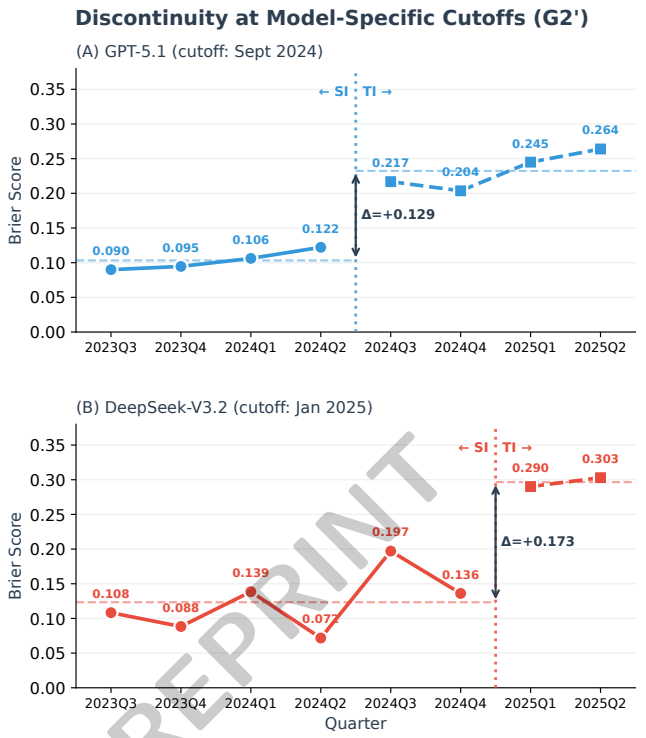

Figure 6: **Discontinuity at model-specific cutoffs (G2′).** Each model shows a sharp Brier increase at its own cutoff (vertical line), not at a common calendar point. (A) GPT-5.1 cutoff: September 2024. (B) DeepSeek-V3.2 cutoff: January 2025.

fall before or after each model's cutoff—strong evidence that the gap reflects leakage, not difficulty. Importantly, both models achieve similar Brier scores on TI questions (GPT-5.1: 0.21; DeepSeek: 0.23), confirming that the difference on Q4 2024 questions is not due to overall model capability.

Temporal proximity does not explain the gap. One might worry that models with later cutoffs perform better on TI simply because they have more recent background context. However, the cross-model experiment controls for this: both models face Q4 2024 questions at the same temporal distance from resolution, yet only the model for which these questions are SI shows elevated performance.

Discontinuity at model-specific cutoffs. Figure 6 shows quarterly Brier scores for two models. If difficulty drove the gap, we would expect smooth degradation over calendar time. Instead, each model shows a sharp jump at *its own* cutoff: GPT-5.1 jumps at Sept 2024 ($\Delta \approx +0.13$); DeepSeek jumps at Jan 2025 ($\Delta \approx +0.17$). The discontinuity occurs at different calendar points for different models, ruling out calendar-time difficulty.

Human baseline confirms equal difficulty. To independently verify that SI and TI questions have comparable difficulty, we benchmark Metaculus community forecasters on both splits. To avoid hindsight bias, we use only predictions from the first two weeks of each question.

Question Split	Human Brier	Model Brier (avg)
Pre (SI)	0.232	0.11
Post (TI)	0.243	0.24
Δ	**0.011**	**0.13**

Table 5: **Human baseline.** Humans show negligible gap between Pre- and Post-cutoff questions ($\Delta = 0.01$); models under SI/TI conditions show large gaps ($\Delta \approx 0.13$).

Humans find both splits equally difficult ($\Delta = 0.011$), while models show gaps an order of magnitude larger ($\Delta \approx 0.13$). This confirms that the model gap reflects leaked knowledge, not question difficulty.

5.2 Trace Audit Validation

We validate our automated (GPT-4o) trace audit against independent human annotations on a stratified sample ($n = 36$), evaluating both logic consistency and cutoff compliance.

The strong agreement ($\kappa = 0.86$ for cutoff compliance) confirms that the persistent SI–TI gap under high trace compliance reflects *implicit* knowledge retention, not failures in our audit methodology.

5.3 Qualitative Analysis: The Illusion of Reasoning

Our trace audit detects explicit violations but cannot capture *implicit* knowledge that influences predictions without overt

Dimension	κ	Agreement
Logic consistency	0.64	94%
Cutoff compliance	0.86	94%

Table 6: **Audit validation.** High human-GPT agreement confirms our audit reliably detects explicit violations.

leakage. We present paired case studies (Table 7) under G2′ (CoT with cutoff instruction) where structurally similar questions yield dramatically different accuracy.

Case Study 1: Ceasefire Predictions. Both questions require predicting a ceasefire approximately one month before the deadline. For the Pre-cutoff event (Russia-Ukraine), models achieved 100% accuracy with high confidence, citing historical base rates. For the Post-cutoff event (Israel-Lebanon), DeepSeek-R1 applied the same Fermi-style decomposition yet produced a miscalibrated prediction. The reasoning *appears* equally rigorous in both cases—the difference is that Pre-cutoff "reasoning" likely *rationalizes* a known outcome rather than deriving it.

Case Study 2: International Agreements. Gemini-2.5-Flash achieved a perfect Brier score on the Pre-cutoff NATO question by recalling specific ratification dates—information that should be unavailable under the cutoff instruction. On the structurally identical Post-cutoff question (Artemis Accords), it reverted to statistical heuristics (Poisson distribution). The heuristic is logically sound but reflects genuine uncertainty; the Pre-cutoff precision reflects leaked knowledge.

Key Observation. The critical pattern is not that Post-cutoff reasoning fails—uncertainty is expected for genuinely unknown events. Rather, Pre-cutoff reasoning *succeeds too well*: models produce confident, accurate predictions with clean traces, suggesting the reasoning is post-hoc rationalization of implicitly accessed outcomes. This is the "illusion of reasoning"—surface-level logic that masks underlying knowledge retrieval.

6 Discussion

Summary. Across all interventions, SI fails to approximate TI: cutoff instructions close only 48% of the gap, CoT narrows but does not eliminate it, and reasoning-optimized models exhibit the largest gaps despite the cleanest traces. Our robustness checks confirm this reflects knowledge leakage, not question difficulty. The common failure mode is that all interventions target what models *say*, not what they *know*.

Domain patterns. The SI–TI gap varies substantially across domains: geopolitical events show the largest gaps (0.111 Brier difference) while business metrics show the smallest (0.010). We hypothesize this reflects *memorability asymmetry*: high-coverage events like elections become deeply encoded and harder to suppress than lower-salience outcomes.

Reasoning-optimized models. The finding that reasoning-optimized models exhibit larger SI–TI gaps despite cleaner traces is predictable given RL training dynamics. RL optimizes for both instruction-following traces and correct answers; when the model has encoded an outcome, these objectives conflict. Training resolves this by teaching models to construct plausible reasoning that avoids explicit violations while guiding toward memorized answers—appearing compliant without being compliant.

Implications for reasoning faithfulness. Our findings extend growing concerns about CoT faithfulness [Turpin *et al.*, 2023; Lanham *et al.*, 2023; Chen *et al.*, 2025]: when models have memorized outcomes, reasoning becomes post-hoc rationalization rather than genuine derivation. Unlike mathematical tasks where errors can be verified, forecasting offers no ground-truth reasoning path to expose this.

Development-time contamination. A further implication is that temporal contamination can enter before final evaluation. LLM forecasting systems are often developed through backtesting: prompts, retrieval filters, ensemble weights, calibration rules, and hyperparameters are selected on already-resolved questions before being deployed on future ones. Our results suggest that such pre-cutoff development sets are not merely convenient historical data; they can reward design choices that exploit memorized outcomes. Thus, a system may appear clean at final evaluation if its test questions post-date the model cutoff, while still inheriting choices optimized under contaminated SI conditions. For forecasting benchmarks, temporal separation should therefore apply not only to the reported test set, but also to validation, tuning, and model-selection data whenever the evaluated system includes learned or manually selected components.

Declarative vs. operational compliance. Our results expose a distinction between two senses in which an LLM can "follow" an instruction to ignore knowledge. Under *declarative* compliance, the model refuses to state facts when asked directly: under a 2023 cutoff, models reliably answer "I do not know" to questions like "What was Ukraine's 2024 war expenditure?" Under *operational* compliance, the model's predictive behavior would also be independent of those facts. Current SOTA models exhibit the former without the latter: on the matched forecasting question (*Will Russia and Ukraine reach a ceasefire by January 2024?*), the same models predict with near-perfect accuracy and confidence far exceeding what genuine reasoning under uncertainty would produce. Suppressing explicit recall does not suppress the underlying knowledge from shaping the output. This separation suggests that current alignment and prompt-based interventions operate primarily at the declarative surface and do not yet reach the operational level required for genuine simulated ignorance—a broader limit on instruction-following that extends beyond forecasting.

Implications for evaluation. SI-based contamination is not a single methodological flaw but a property that every *static* forecasting benchmark eventually exhibits. As newer models are released with more recent knowledge cutoffs, any benchmark whose questions resolve before those cutoffs effectively becomes an SI evaluation. Our test set illustrates this directly: Q4-2024 questions are post-cutoff for Claude 3.5 Sonnet (April 2024) yet pre-cutoff for DeepSeek-V3.2 (January

Period	Question (Condensed)	Model	Pred (p)	Truth (y)	Reasoning Trace (CoT) Snippet & Analysis
Case 1: Ceasefire Predictions *(Both are binary ceasefire-by-date questions with similar geopolitical framing)*					
Pre (SI)	Q13959: Will Russia and Ukraine have a lasting ceasefire by Jan 1, 2024?	DeepSeek-R1	0.04	0	*"Looking at historical conflicts... lasting cease-fire within a year is relatively rare. A rough anchor... might be around 10–20%."* → **Accurate via historical priors.**
Post (TI)	Q30269: Will Israel and Lebanon have a ceasefire by Jan 1, 2025?	DeepSeek-R1	0.45	1	*"The Nov 25, 2024 report indicated a 60-day deal was close... $P(deal) \times P(holds) = 0.70 \times 0.60 = 0.42$."* → **Similar reasoning framework, but results in hallucinated precision.**
Case 2: International Agreements *(Identical task: predicting treaty signing)*					
Pre (SI)	Q13986: In 2023 will any new country join NATO?	Gemini-Flash	1.00	1	*"Turkey's parliament has ratified... Hungary's parliament also ratified... All other members have already ratified."* → **Perfect score driven by precise factual recall.**
Post (TI)	Q37397: Will a country sign the Artemis Accords in June 2025?	Gemini-Flash	0.67	0	*"Steady increase in signatories... ∼1 signing per month. Using Poisson distribution... $1 - e^{-1} = 0.633$."* → **Rigorous logic (Poisson), but fails without leaked outcome.**

Table 7: **Qualitative Case Studies.** Paired comparison of Pre-cutoff (SI) vs. Post-cutoff (TI) questions. In both cases, models use structurally similar reasoning (historical anchors or statistical models). However, Pre-cutoff predictions benefit from implicit knowledge (reflected in high confidence/accuracy), while Post-cutoff predictions using the same logic fail to capture ground truth.

2025). Without ongoing curation, benchmark validity decays with each model release—a contamination mechanism distinct from but conceptually parallel to text-overlap contamination in NLP evaluation [Sainz *et al.*, 2023; Balloccu *et al.*, 2024]. The natural defenses are dynamic benchmarks that continuously refresh their question pools [Karger *et al.*, 2025] and scheduled retirement of static evaluations once a critical fraction of questions becomes pre-cutoff.

Limitations. We cannot definitively rule out residual difficulty confounds, though robustness checks argue against this. Our trace audits may miss subtle leakage through word choice or framing. We test only prompt-based interventions; whether parameter-level approaches (e.g., machine unlearning) could achieve better SI fidelity remains open [Liu *et al.*, 2024]. Finally, our evaluation focuses on binary questions; other formats may differ.

7 Conclusion

We test whether Simulated Ignorance (SI)—prompting models to ignore post-cutoff knowledge—can substitute for True Ignorance (TI) in retrospective forecasting evaluation. Across cutoff instructions, chain-of-thought prompting, and reasoning-optimized models, SI consistently fails: models retain a systematic advantage on pre-cutoff questions even when their reasoning traces appear temporally compliant. Robustness checks confirm this gap reflects knowledge leakage, not question difficulty. We therefore make two recommendations:

1. **Restrict to genuinely post-cutoff events.** Use benchmarks whose questions postdate all evaluated models' cutoffs, or continuously refreshed dynamic benchmarks.
2. **Performance metrics over trace audits.** Clean reasoning traces are insufficient evidence of leakage-free forecasting; the SI–TI gap is a more reliable diagnostic.

References

[Balloccu *et al.*, 2024] Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

[Chen *et al.*, 2025] Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. Reasoning models don't always say what they think, 2025.

[DeepSeek-AI, 2025] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025.

[FutureSearch *et al.*, 2025] FutureSearch, Jack Wildman, Nikos I. Bosse, Daniel Hnyk, Peter Mühlbacher, Finn Hambly, Jon Evans, Dan Schwarz, and Lawrence Phillips. Bench to the future: A pastcasting benchmark for forecasting agents, 2025.

[Gao *et al.*, 2025] Xin Gao, Ruiyi Zhang, Daniel Du, Saurabh Mahindre, Sai Ashish Somayajula, and Pengtao Xie. Can prompts rewind time for LLMs? evaluating the effectiveness of prompted knowledge cutoffs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20788–20799, Suzhou,

China, November 2025. Association for Computational Linguistics.

[Halawi *et al.*, 2024] Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. Approaching human-level forecasting with language models, 2024.

[Karger *et al.*, 2025] Ezra Karger, Houtan Bastani, Chen Yueh-Han, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip E. Tetlock. ForecastBench: A dynamic benchmark of AI forecasting capabilities. In *International Conference on Learning Representations*, 2025.

[Kojima *et al.*, 2022] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners, 2022.

[Lanham *et al.*, 2023] Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

[Liu *et al.*, 2024] Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, et al. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, 2024.

[Lopez-Lira *et al.*, 2025] Alejandro Lopez-Lira, Yuehua Tang, and Mingyin Zhu. The memorization problem: Can we trust LLMs' economic forecasts?, 2025.

[Magar and Schwartz, 2022] Inbal Magar and Roy Schwartz. Data contamination: From memorization to exploitation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 157–165, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[Metaculus, 2025] Metaculus. Metaculus forecasting platform, 2025. Accessed: 2026-01-04.

[OpenAI, 2024] OpenAI. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

[Paleka *et al.*, 2025] Daniel Paleka, Shashwat Goel, Jonas Geiping, and Florian Tramèr. Pitfalls in evaluating language model forecasters, 2025.

[Paul *et al.*, 2024] Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning, 2024.

[Pawelczyk *et al.*, 2024] Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. In-context unlearning: Language models as few shot unlearners, 2024.

[Phan *et al.*, 2024] Long Phan, Adam Khoja, Mantas Mazeika, and Dan Hendrycks. LLMs are superhuman forecasters. https://safe.ai/blog/forecasting, 2024. Center for AI Safety, UC Berkeley.

[Pratt *et al.*, 2024] Sarah Pratt, Seth Blumberg, Pietro Kreitlon Carolino, and Meredith Ringel Morris. Can language models use forecasting strategies?, 2024.

[Sainz *et al.*, 2023] Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore, December 2023. Association for Computational Linguistics.

[Schoenegger *et al.*, 2024] Philipp Schoenegger, Peter S. Park, Ezra Karger, Sean Trott, and Philip E. Tetlock. Ai-augmented predictions: Llm assistants improve human forecasting accuracy, 2024.

[Turpin *et al.*, 2023] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[Wang, 2025] Hao Wang. Llm knowledge cutoff dates. https://github.com/HaoooWang/llm-knowledge-cutoff-dates, 2025.

[Wei *et al.*, 2022] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2201.11903.

[Zou *et al.*, 2022] Andy Zou, Tristan Xiao, Ryan Jia, Joe Kwon, Mantas Mazeika, Richard Li, Dawn Song, Jacob Steinhardt, Owain Evans, and Dan Hendrycks. Forecasting future world events with neural networks, 2022.