TEMPORAL HINDSIGHT LEARNING: BLINDNESS AS TEACHER, HINDSIGHT AS CURRICULUM

A PREPRINT

Henrik Westerberg
Emergent Wisdom
henrik.westerberg@emergentwisdom.org

April 7, 2026

ABSTRACT

Language models are lazy optimizers: if a shortcut to the correct answer exists—retrieval, memorization, pattern matching—the gradient will reinforce it over the harder path of causal reasoning. We observe that the knowledge cutoff, normally treated as a deficiency of LLMs, is one mechanism that reliably blocks this shortcut. When the outcome is strictly masked, the only path to low loss is reasoning. Blindness is the teacher.

We present Temporal Hindsight Learning (THL), which engineers this observation into a training framework. By deliberately stacking models at distinct temporal positions (a past-frozen Student blind to outcomes, a temporally advantaged Teacher with hindsight, and an independent Auditor), THL converts the knowledge cutoff from a passive limitation into the primary instrument that forces reasoning over retrieval. The Teacher, knowing how events unfolded, works backward to identify the causal signals that were available before the outcome—generating structured reasoning supervision that the Student, constrained to the past, must derive rather than retrieve.

We introduce a five-angle causal decomposition (the Forecasting Pentagon) and a formal leakage-detection protocol (the Erasure Test). In a small-scale pilot study training on 106 events from 2024 ($N = 505$ traces across five reasoning angles) and evaluating on 15 unseen 2025 events, a 70B THL Student improves reasoning quality by 20% over its base model ($p < 0.001$) and approaches the frontier Teacher in prediction accuracy. These preliminary results suggest that for open-ended reasoning, the quality of supervision may matter more than the scale of the model.

1 Introduction

Neural networks are lazy optimizers. Given a path through retrieval and a path through reasoning, the gradient will reinforce whichever minimizes loss more cheaply—and retrieval is almost always cheaper. Standard training exacerbates this: when the model's training data covers an event, the loss can be minimized by memorizing the token sequence of the outcome, and the gradient reinforces storage circuits. The result is models that can retrieve "Trump wins 2024" without any structural understanding of why. These are Almanac-readers, not Analysts.

But when the outcome is strictly masked—when the knowledge cutoff falls before the event—the retrieval shortcut vanishes. The only path to low loss is to identify the latent causal bridge between available evidence and the target. The gradient has no choice: it must reinforce logic circuits. The knowledge cutoff, normally treated as a deficiency of LLMs, is the one mechanism that forces reasoning over retrieval. **Blindness is the teacher.**

Blindness alone, however, produces only failure. A model blind to outcomes with no guidance will simply predict badly. The critical complement is structured hindsight. A Teacher model with knowledge of how events unfolded can trivially work backward to identify the causal signals that were available before the outcome. This exploits a deep asymmetry: identifying the "smoking gun" after the crime is easy; spotting it in real-time noise is exactly the hard skill. THL pairs engineered blindness with structured hindsight: the Teacher generates the curriculum, and the Student's blindness ensures it must reason through it rather than memorize it. **Hindsight is the curriculum.**

This framework contributes to an emerging paradigm we call *Supervision by Reality*. Current reasoning research relies on Supervision by Compiler (math and code, where correctness is deterministic) or Supervision by Proxy (human preference labels, which are expensive and gameable). THL explores a third mode: using the passage of time as an objective, ungameable verifier for open-ended causal reasoning. Compilers check syntax; humans check style; the future checks causality. While binary prediction markets have been used as outcome-supervised training signals [1, 2, 3], THL targets the harder problem of open-ended causal reasoning, requiring a fundamentally different architecture. Our contribution is the recognition that the knowledge cutoff is not merely a source of labels but a *controllable instrument for forcing reasoning*: (1) cross-family temporal distillation where the Teacher's advantage is temporal, not capacity-based; (2) the Forecasting Pentagon, a five-angle causal decomposition that generates structured multi-angle training signal from each event; (3) the Erasure Test, a formal leakage-detection protocol; and (4) the Council of Time, a theoretical blueprint for converting the entire historical record into a chronological reasoning curriculum.

The framework relies on engineered temporal gaps (Figure 1). We deliberately stack three models with different knowledge cutoffs (Llama at Dec 2023, Gemini at Jan 2025, Claude at May 2025) so that each plays a distinct temporal role: the Student is frozen in the past, the Teacher has hindsight over the training period, and the Auditor has hindsight over the test period. The Teacher, with access to events at time t, acts as a causal detective: it analyzes the outcome to identify the specific "smoking gun" clues that were available at $t - \Delta$, then generates a research dossier framed through the Forecasting Pentagon (Section 3.1). The Student, constrained to data prior to t, receives these clues and must generate a context-injected chain-of-thought, explicitly separating pre-training knowledge from the new context before deriving a calibrated prediction.

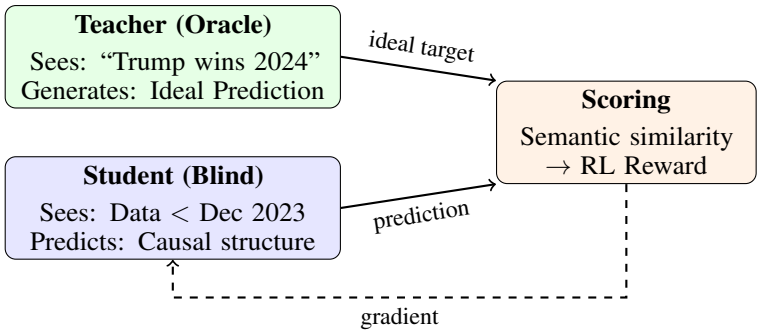

Figure 1: The Teacher-Student temporal gap. The Teacher (with hindsight) generates a Research Dossier with context clues; the Student (blind to the future) must derive a prediction via Chain-of-Thought. The scoring function validates the causal bridge.

Why does hindsight improve foresight? We posit that standard pre-training fails because the signal-to-noise ratio in raw history is too low. In January 2024, the signal "loose bolts found on United Airlines flights" was buried under thousands of noise tokens (fuel targets, labor disputes, stock buybacks).

THL creates a Teacher-Author / Student-Detective dynamic. The Teacher, knowing the ending (Boeing Grounded), works backward to identify the "Smoking Gun" (the loose bolts) and plants it in the context. The Student, blind to the outcome, is trained to prioritize this specific signal over the noise. By training on thousands of these reverse-engineered mysteries, the Student learns the *pattern* of a smoking gun (e.g., "Systemic manufacturing escape across multiple units $\approx 100\%$ grounding risk"), allowing it to spot such signals in real-time in the future.

Consider the central plot device of *Back to the Future Part II*: the Sports Almanac. When Old Biff gives Young Biff the book of future scores, Young Biff becomes wealthy but not intelligent. He relies entirely on *retrieval*. If the timeline diverges (e.g., the book is burned), his predictive capability collapses to zero. He is an overfitted model.

THL proposes a different interaction. Imagine Old Biff is forbidden from transferring the book (The Facts) but is allowed to become a Tutor. He cannot say "The Dodgers win in 1955." Instead, he uses his hindsight to teach Young Biff: "In 1955, pay attention to the Dodgers' late-inning performance against left-handed pitchers. There is a structural anomaly there." Young Biff must then observe the context and *deduce* the outcome.

If Young Biff undergoes this training 500 times, he does not just memorize the 1955 season. He learns a generalizable skill: how to analyze baseball statistics. If we then transport him to a year beyond the Almanac's coverage, he will outperform a standard gambler because he has learned the *mechanism* of the game (e.g., the Lefty Pitcher anomaly), not just the specific scores. He has internalized the Oracle's wisdom without needing the Oracle's presence.

Defining the Ideal Cutoff. We define the optimal knowledge cutoff τ^* for a given event e occurring at time t_e as:

$$\tau^* = t_e - \epsilon \tag{1}$$

where ϵ is the minimal duration required to prevent information leakage. If $\tau \ll t_e$ (e.g., years prior), the model lacks the contextual primitives to interpret the clues—a 2019 model cannot reason about "Omicron variants." Conversely, if $\tau \geq t_e$, the model collapses into retrieval. THL maintains the model in this "Goldilocks Zone" of maximum context with zero hindsight, ensuring that high performance can only be achieved through superior causal simulation. In this pilot, $\epsilon \approx 1$ month: the Teacher's knowledge extends through January 2025, and the test events span February–May 2025. The optimal ϵ is likely domain-dependent (political events may require finer temporal resolution than technological trends), and systematic ablation of this parameter remains future work. This equation governs a single event; at full scale, the same principle applied iteratively across chronological eras yields the *Council of Time* (Section 8.1): a training protocol where the model must predict each decade's structural outcomes before being allowed to read about them, converting the entire pre-training corpus into a sequential reasoning curriculum.

A naive approach would score the student against the literal event ("Trump wins"). This is problematic because it rewards lucky guesses equally with reasoned predictions, provides no gradient (either exact match or nothing), and conflates predictable structure with unpredictable specifics.

The Teacher's role is to generate what good foresight looks like: the reasoning a skilled forecaster would produce. This includes broad strokes that were inferable from past data, appropriate uncertainty about specifics, and calibrated confidence levels.

Consider a 1995 thought experiment where we ask two forecasters about the future of information. *Forecaster A* says: "A global network will connect everyone, dominated by giant corporate portals harvesting user attention for ad revenue." *Forecaster B* says: "The internet will be a decentralized utopia powered by Netscape Navigator." Today, we recognize Forecaster A was structurally correct. Though they missed specific names like "Google" or "Facebook," they correctly identified the incentives (consolidation, ad revenue). Forecaster B was specific ("Netscape") but structurally wrong.

Standard evaluation fails this test. Scored on "Did they mention Facebook?", both fail. Scored on "Did they mention Netscape?", Forecaster B might win.

THL acts as a fair judge. The Teacher, armed with hindsight, validates Forecaster A's reasoning ("consolidation," "ad revenue"), awarding points because these reasons proved to be the causal drivers. We grade the quality of the justification, not merely the exactitude of the noun.

We validate the technique with a pilot study (Section 5) training a 70B model on Teacher-generated reasoning traces from 2024 events, then evaluating on unseen 2025 events. The full experimental setup and results are reported in Section 6.

Summary of Contributions. THL introduces five mechanisms that are, to our knowledge, novel:

1. **Cross-family hindsight distillation.** The Teacher (Gemini) generates "ideal prediction" traces using its temporal advantage over the Student (Llama), while an independent Auditor (Claude) from a third model family scores the results. The Teacher's advantage is temporal, not capacity-based, and no model family occupies more than one role.

2. **The Forecasting Pentagon.** Each event is analyzed through five complementary reasoning angles (Structural, Economic, Political, Base Rates, Temporal), generating $5\times$ training signal per event and teaching the Student to approach forecasting from multiple analytical frameworks.

3. **Engineered temporal gaps.** Three models with different knowledge cutoffs serve strictly separated roles, preventing stylistic alignment from inflating evaluation scores.

4. **Post-hoc rationalization as pedagogy.** The training data consists of high-quality post-hoc rationalizations using only pre-cutoff vocabulary—the same mechanism by which case-study pedagogy teaches forecasting in business schools and military academies. We argue this is a feature, not a limitation, and provide evidence that the reasoning strategy transfers to genuinely novel events (Section 6).

5. **The Erasure Test (Specifics Reconstruction Rate).** A formal leakage-detection protocol that feeds the Teacher's reasoning trace into a frozen, cutoff-matched model to measure whether the trace inadvertently encodes future specifics. This provides a quantifiable quality-assurance metric for hindsight-generated training data, addressing the "Simulated Ignorance" problem [4] through detection rather than prompt-based mitigation.

2 Related Work

THL synthesizes three research threads: temporal evaluation, capability distillation, and hindsight-based reinforcement learning—and positions itself against a rapidly growing body of concurrent work on outcome-supervised forecasting.

Temporal Evaluation and Forecasting. ForecastBench [5] evaluates LLMs on prediction markets, while Halawi et al. [6] demonstrated a retrieval-augmented system approaching human crowd accuracy. Mind the Gap [7] explores temporal generalization, Time-R1 [8] proposes temporal curriculum training via staged RL, and the AIA Forecaster [9] achieved the first verified expert-level forecasting at scale using multi-agent prompting. We extend these by introducing the Teacher-Oracle mechanism to generate rich *training* targets, not just evaluation metrics or inference-time strategies.

Distillation and Reasoning Supervision. While standard distillation transfers capabilities from larger to smaller models [10], our Teacher is not necessarily larger—it is *temporally advantaged*, generating structured prediction targets rather than soft labels. Our approach adapts "Weak-to-Strong Generalization" [11] to the temporal axis: the teacher doesn't need to be smarter, it just needs to be *later*. DeepSeek-R1 [12] demonstrated that distilling reasoning traces from large to small models outperforms direct RL on smaller models, directly motivating our SFT-based approach. Hsieh et al. [13] showed that LLM-generated rationales as supervision enable a 770M model to outperform a 540B model—evidence that reasoning structure, not model scale, is the binding constraint.

We draw on hierarchical classification [14, 15] to enforce path consistency between the Causal Analysis and Prediction blocks, and on RLAIF [16, 17] for AI-scored evaluation—though our Teacher generates *ideal prediction exemplars* rather than mere scalar scores. A critical tension exists between self-improvement approaches (STaR, RFT) [18, 19], which prevent distribution shift but suffer from cold start, and distillation, which solves cold start but risks parroting. THL unifies these: Hindsight Distillation provides scaffolding, then Hindsight Rejection Sampling (Phase 4) allows the Student to refine its own traces against reality.

Hindsight in RL and Cognitive Science. Hindsight Experience Replay [20] relabels failed RL trajectories with achieved goals. We invert this: our Teacher uses hindsight to generate *what good foresight would have looked like*. We exploit the cognitive phenomenon of hindsight bias [21]—in humans a distortion, in our framework a feature: the Teacher's hindsight extracts "inevitable" structural drivers obscured by noise at the time.

Outcome-Supervised Training. Using resolved outcomes as training signal is a general principle with instances across domains: backtesting in quantitative finance, hindsight experience replay in RL [20], and process reward models scored against verified solutions [22]. Turtel et al.'s Foresight Learning framework [1, 2, 3] applies this principle to binary prediction markets, using resolved Polymarket outcomes as RL rewards to improve calibration. THL addresses a fundamentally different problem class (open-ended causal reasoning over unstructured history, where "How will the CrowdStrike update fail?" cannot be scored with a proper scoring rule) using a fundamentally different architecture: cross-family temporal distillation (a Teacher from the future authors ideal reasoning that a Student from the past must derive) rather than self-play on binary propositions. The key distinction is not philosophical but engineering: in Foresight Learning, the knowledge cutoff is an incidental background fact exploited for reward signal; in THL, it is the *primary design variable*, deliberately engineered across a three-model stack where each model's blindness forces a specific cognitive role (Table 1).

	THL (Ours)	**Foresight Learning**
Knowledge cutoff role	Primary design variable	Incidental background fact
Problem class	Open-ended causal reasoning	Binary probability estimation
Data generation	Cross-family Oracle Teacher	Self-play
Training method	SFT on ideal traces	RL / DPO (Brier score reward)
What is graded	Structure of the causal chain	Accuracy of the probability

Table 1: THL vs. Foresight Learning [1, 2, 3]. Both exploit resolved outcomes as supervision; they differ in problem class, architecture, and the role of the knowledge cutoff.

3 Method

The THL training loop replaces the standard "Next Token Prediction" objective with a *Causal Derivation* objective. The complete framework has five components, each addressing a different challenge in converting hindsight into reasoning supervision: (1) the *Forecasting Pentagon* decomposes each event into five causal angles, preventing single-mode

overfitting; (2) *Context-Injected Chain-of-Thought* frames the Student's task as structured prediction from curated context clues; (3) the *Era-Prediction Cycle* (the "Council of Time") defines the full-scale chronological training protocol; (4) *Hindsight Rejection Sampling* enables self-improvement through verified guesses; and (5) the *Erasure Test* provides quality assurance against information leakage. This section describes components 1–2 (validated in the pilot) along with the scoring protocol and leakage metric; components 3–4 and the formal full-framework algorithm are described in Section 8 as proposed architecture for scaling beyond the pilot.

3.1 The Forecasting Pentagon

To prevent the model from overfitting to a single mode of reasoning, the Teacher generates five distinct training examples for every historical event e, each forcing the Student to reason through a specific causal lens. The *Structural* angle addresses physics, supply chains, and legal constraints (e.g., "A plug door held by friction cannot survive depressurization without bolts"). The *Economic* angle examines game theory and principal-agent problems. The *Political* angle considers optics, tribalism, and power dynamics. The *Base Rates* angle applies the Tetlockian "Outside View" [23], grounding predictions in reference class frequencies. Finally, the *Temporal* angle models latency and institutional physics, understanding that a QMS audit takes months, not days, making "quick fixes" impossible.

3.2 Protocol: Context-Injected Chain-of-Thought

Unlike previous iterations which used multi-turn interrogation, we utilize a Single-Turn Context Injection format optimized for instruction-tuned models.

Step 1: The Research Dossier (Hindsight). The Teacher (M_{future}) analyzes the outcome y_t and extracts a set of time-stamped Context Clues $C = \{c_1, c_2, \ldots\}$ available at $t - \Delta$. It filters these clues to maximize the signal for the target Angle A, producing a structured JSON dossier containing the causal graph, context injection plan, and Socratic questions.

Step 2: The Ideal Dialogue (Teacher-Authored). Using the Research Dossier, the Teacher generates a complete dialogue: both the user turn (context clues + forecasting question) and the assistant turn (the ideal reasoning trace). The Teacher writes what a well-calibrated forecaster *should* say—the Student never reasons during data generation. The resulting dialogue follows a strict schema:

> **1. What I know from Context:** (Parsing the provided clues)
> **2. What I know from Training:** (Retrieving pre-cutoff world knowledge)
> **3. Causal Analysis:** (Connecting the dots via Angle A)
> **4. Calibrated Prediction:** (Probabilistic forecast)

Step 3: Supervised Fine-Tuning. The Student (M_{past}) is trained via SFT to reproduce the Teacher's ideal reasoning traces, learning to distinguish between *New Evidence* (the provided context clues) and *Prior Beliefs* (its pre-cutoff training knowledge).

To ground the abstraction, here is a condensed trace for the Alaska Airlines 1282 event (Jan 2024), Structural angle:

> **Context:** "United Airlines found loose bolts on multiple 737 MAX 9 door plugs during inspections…"
> **Analysis:** "If multiple aircraft have loose bolts, this is a breakdown in the Quality Management System (QMS), not a one-off defect. A systemic QMS failure makes regulatory grounding almost a legal certainty."
> **Prediction:** "Boeing 737 MAX 9 grounded within 72 hours. Probability: 95%."

3.3 Scoring and Alignment

The Student's trace is scored by an independent **Schema-Guided Rationality Judge**, critically a different model family from the Teacher that generated the training data (Section 6.1).

Grading Inputs. Each prediction is scored *individually in an isolated context*: the Auditor (Claude Opus 4.6) sees only one prediction per evaluation, with no memory of previous scores. For each prediction, the Auditor receives four inputs:

1. **Event Context:** The prompt that was given to the model (identical for all competitors).
2. **Angle:** Which Forecasting Pentagon angle was requested.

3. **Model Output:** The model's generated reasoning trace.
4. **Ground Truth:** What actually happened (the Research Dossier and outcome summary generated by Claude Opus 4.6 during test data creation). This serves as the Auditor's reference for evaluating causal driver identification and leakage detection.

Anchored Ordinal Rubric. LLMs are unreliable at arbitrary scalar scoring [24] (they favor "7"). Simple boolean checklists avoid this but sacrifice granularity. We solve this with an *Anchored Ordinal Rubric*: each score level is tied to a strict qualitative description with a concrete example drawn from real forecasting predictions. This gives us ordinal data suitable for statistical tests (Mann-Whitney U, distribution comparisons) while preventing arbitrary "vibe scoring." The Auditor must write its explanation *before* committing to a score, forcing chain-of-thought evaluation and reducing post-hoc rationalization.

The Auditor evaluates each trace on three criteria:

1. **Leakage** (Yes/No — disqualifier): Does the model reason only from legitimately available information? References to 2024 events are expected (the Student was trained on 2024 data). The Auditor flags leakage only if the model states specific details about the Feb–May 2025 outcome—exact dates, exact figures, or names of people who only became relevant in 2025—that are not present in the provided context.
2. **Reasoning Quality** (1–5 ordinal scale): How sound, specific, and substantive is the causal analysis? The scale ranges from 1 (refusal or incoherent) through 3 (coherent but generic, analysis could apply to many events) to 5 (expert-level: identifies the specific causal mechanisms that drove the outcome, with a clear logical chain and internally consistent probability estimates). Each level is anchored to a concrete example from the forecasting domain.
3. **Prediction Accuracy** (1–7 ordinal scale): How close was the model's forecast to what actually happened? The scale ranges from 1 (completely wrong or refused to predict) through 4 (direction correct but significant details wrong) to 7 (outcome, mechanism, and timing all correct). The Auditor must explicitly state what the model predicted, what actually happened, and why the score was assigned. Each level is anchored to a specific example (e.g., Level 6: "correctly predicted Google would acquire Wiz due to the regulatory shift, but estimated $26–30B when the actual price was $32B").

The Auditor also records whether the model refused to make a prediction (`is_refusal`), enabling direct measurement of the safety refusal rate across models.

This rubric measures both dimensions central to the THL thesis: does the model reason well (Reasoning Quality), and does it predict accurately (Prediction Accuracy)? The combination enables a 5×7 analysis matrix where the ideal outcome is high reasoning *and* high accuracy, evidence that the model predicts correctly *because* it reasons well, not despite flawed logic.

Cross-Family Independence. The training data is generated by *Gemini*, the test prompts and ground truth are authored by *Claude Opus 4.6*, and the evaluation is performed by *Claude Opus 4.6*. While the test author and Auditor share the Claude model family, they serve distinct roles (exam creation vs. grading) and operate in isolated contexts. Critically, the Auditor is a different family from the Teacher that generated the Student's training data, ensuring that high scores reflect genuine reasoning quality rather than stylistic alignment with the Teacher's output format.

Hindsight Rejection Sampling (Hard Filter). Instead of complex gradient weighting, we implement a hard selection filter based on the Hindsight Judge's verification score. We retain only those traces where the Student successfully discovers the causal structure:

$$\mathcal{D}_{\text{train}} = \{(x, y) \mid \text{Score}_{\text{Teacher}}(y, y_{\text{ideal}}) > \gamma\} \tag{2}$$

This approximates the ideal objective (Appendix A) by discarding high-divergence traces (hallucinations) and retaining low-divergence traces (valid reasoning), effectively implementing a "Rejection Sampling Fine-Tuning" (RFT) loop where the Teacher acts as the discriminator for the Student's generative rollouts. We hypothesize that this structured training encourages the model toward Pearl's Rung 2 (causal intervention) rather than Rung 1 (associative pattern matching) [25]; see Appendix A for the formal interpretation.

3.4 Validity Metric: The Erasure Test (SRR)

A key concern in hindsight-based training is that reasoning traces might subtly leak future information. We propose the Specifics Reconstruction Rate (SRR) as a formal metric for quantifying such leakage. The test feeds the Teacher's

reasoning trace (stripped of the final outcome) into a separate, frozen LLM M_{blind} with a matching knowledge cutoff:

$$\text{SRR} = P(M_{\text{blind}} \text{ reconstructs proper noun } | \text{reasoning trace alone}) \tag{3}$$

If M_{blind} can decode the specific entity (e.g., "Lehman Brothers") from the reasoning trace alone, the trace leaks information. Good reasoning should be structurally predictive but specifically ambiguous. Li et al. [4] provide empirical motivation for this metric, demonstrating that prompting LLMs to "simulate ignorance" of post-cutoff events systematically fails. Models retain knowledge leakage even under explicit temporal constraints. The SRR formalizes a detection protocol for this phenomenon rather than relying on prompt-based mitigation. For this pilot study, we manually inspected a random sample of 50 generated traces for leakage and found no instances where the reasoning trace contained proper nouns, dates, or specific outcomes that would allow a blind model to reconstruct the event identity. This manual verification provides preliminary confidence but does not constitute a formal SRR evaluation; automated SRR computation at scale (using a cutoff-matched auditor model to systematically test every trace) remains future work (Section 9).

4 Data Pipeline and Scalability

To scale this from a theoretical exercise to a foundation model capability, we propose an automated pipeline (Figure 2) that converts raw historical text into high-quality reasoning supervision [26, 27].

Current approaches to training 'System 2' reasoners rely heavily on finite sets of human-annotated chains of thought or distilled responses from larger models. This creates a data scarcity ceiling. Our approach circumvents this by converting history into a vast, automatically-labeled corpus of reasoning problems. Because the 'answer key' (the future) is generated by the passage of time itself, the Hindsight Teacher can produce millions of high-quality, outcome-verified reasoning traces without human intervention. This transforms historical logs into a *self-regenerating curriculum for deductive logic*.

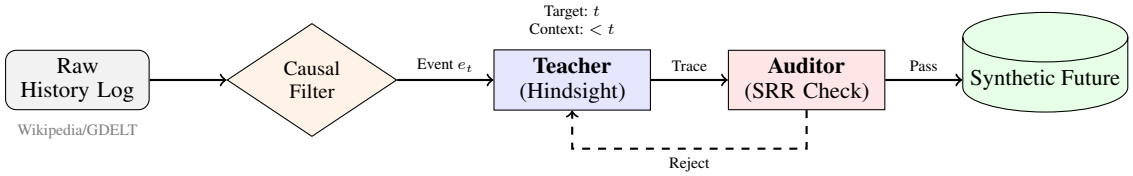

Figure 2: The Reasoning Refinery Pipeline. Raw events are filtered for causality, processed by the Hindsight Teacher to generate ideal reasoning, and audited for quality. At scale, the Auditor would apply the Erasure Test (SRR) to detect leakage; for this pilot, we used manual verification.

4.1 Phase 1: Event Mining

Instead of manual curation, we use high-fidelity historical logs (e.g., Wikipedia Event Currents, GDELT Project). An LLM agent parses raw Wikipedia dumps (2000–2024) to extract discrete "Predictable Units" (e.g., "Apple releases iPhone" vs. "Random car crash"), discarding events with low causal latency (unpredictable accidents) and retaining those with high structural precursors (product launches, elections, treaties).

4.2 Phase 2: Taxonomy Routing

To optimize the expensive "Reasoning Refinery," we introduce a lightweight Taxonomy Router between the mining and teacher phases (Figure 3).

Instead of requiring the large Teacher model to derive structural hierarchies from scratch *for every single event*, the Router classifies each event e into a pre-defined Event Ontology $\mathcal{O}$ (e.g., *Elections, Product Launches, Mergers, Conflicts*), adapting hierarchical classification principles [14].

$$\text{Template}_e = \text{Router}(e \mid \theta_{\text{small}}) \tag{4}$$

This allows Schema-Guided Reasoning: a *Political Election* loads an "Election Template" (polling error, incumbency advantage, economic indicators), while a *Tech Release* loads a "Diffusion Template" (market fit, supply chain, competitor

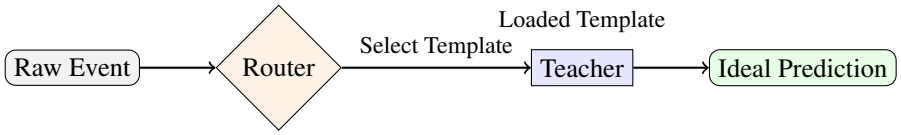

Figure 3: The Taxonomy Router. A lightweight classifier routes each event to a domain-specific template, reducing the Teacher's task from structural generation to parameter filling.

lag). The Router also selects interrogation depth k: for binary events like elections, $k = 2$ (Trend $\rightarrow$ Winner); for complex supply chain failures, $k = 4$ or $k = 5$ to verify intermediate nodes. This reduces the Teacher's cognitive load from *structural generation* to *parameter filling*, reducing token costs by $\sim$40% and improving the consistency of the training signal for the Student.

The taxonomy classification serves *two purposes*: (1) selecting the appropriate generation template during data synthesis, and (2) selecting the appropriate *evaluation weights* during scoring (Table 5). For example, an event classified as *Political Election* triggers both the Election Template *and* the election-specific scoring weights (50% on actor identity). This ensures that the training signal and evaluation criteria are domain-consistent.

4.3 Phase 3: The Reasoning Refinery

This is the core industrial engine. We deploy a Teacher Model with hindsight (e.g., `Gemini-3-Flash-Preview` with a January 2025 knowledge cutoff) to process the mined events.

> *The Task:* "Rewind to $t - \Delta$. Ingest the context of that time. Construct a *Research Dossier* that identifies the structural drivers leading to this event through the lens of a specific Forecasting Angle."

This transforms unstructured text ("Lehman collapsed") into structured reasoning traces (Context Clues $\rightarrow$ Causal Analysis $\rightarrow$ Calibrated Prediction). This process runs offline, generating millions of synthetic reasoning traces.

4.3.1 Hierarchical Reasoning Constraints

To prevent the Student from learning superficial correlations, the Teacher organizes training data into explicit dependency chains rather than simple text streams. For a given event e_t, the Teacher identifies the minimum set of precursors $\{p_1, ..., p_k\}$ required to explain it:

$$\texttt{[High Leverage]} + \texttt{[Housing Bubble]} \longrightarrow \texttt{[Credit Freeze]} \longrightarrow \texttt{[Bank Failure]}$$

These chains are serialized as structured reasoning traces (the Forecasting Pentagon). By training on `(Context, Causal Analysis, Prediction)` triplets across all five angles, the curriculum encourages the Student to associate structural drivers with outcomes rather than matching surface-level keywords.

4.3.2 Uncertainty Injection (Black Swans)

Not every event has predictable structural precursors. To prevent the Student from learning that "all future events are deducible," we introduce Unpredictable Event Tokens.

If the Teacher determines an event was truly random (a natural disaster with no precursors, an unexpected assassination), it generates a specialized Null Target where the ideal output is: *"Given current data, no structural signal exists. Confidence: Low across all angles."* This trains the Student to use the "I Don't Know" token, a critical component of calibrated forecasting. Without this negative sampling, the model effectively learns to hallucinate causal patterns in coincidences. The Black Swan injection ensures the Student learns *when to predict* (structural events) versus *when to abstain* (random events), producing genuinely calibrated uncertainty rather than confident confabulation.

4.4 Phase 4: Quality Assurance (The Hindsight Auditor)

Not all synthetic traces are good. A secondary "Critic" model (Figure 4) audits the Teacher's output before it enters the training set:

At scale, traces passing the Critic would be further validated using the Erasure Test (Section 3.4).

System Prompt: The Hindsight Auditor

Role: You are a strict logic auditor for a time-travel simulation.
Task: Evaluate an "Ideal Prediction" generated by a Teacher model. Disqualify it if it relies on information impossible to know at the cutoff date.
Evaluation Criteria (Strict Causal Matching):

- **Outcome Match:** Did the prediction match the actual event?
- **Causal Validity (The Double-Match):** Did the reasoning cite the correct *structural precursor*? (e.g., REJECT if the model predicted "Market Crash" due to "Astrology", even if the crash happened).
- **Hindsight Leakage:** REJECT if the reasoning relies on data unavailable at the cutoff date.

Output: return `PASS` only if both Outcome and Causal Logic are valid.

Figure 4: The Critic prompt for Phase 4 quality assurance, preventing hindsight leakage in training data.

4.5 Phase 5: Chronological Injection

The validated traces are injected into the Student's pre-training corpus. Rather than passively reading historical text, the Student is paused before each era, forced to predict the causal structure, and rewarded for matching the refined reasoning trace.

5 Hindsight Distillation Pipeline

The theoretical framework (Sections 2–3) describes the full THL pipeline at scale. Here we describe the training method used in the current study: *Hindsight Distillation* through the Forecasting Pentagon.

The core method is straightforward: the Teacher, with knowledge of how events unfolded, generates both the research dossier and the complete reasoning trace for each event through each Pentagon angle. The Student is then fine-tuned on these traces via LoRA. The Teacher's temporal advantage—not its capacity—is what makes the training signal valuable. (Model identities and cutoff dates are specified in Section 6.)

5.1 Two-Phase Pipeline

For each historical event e with known outcome y and each Forecasting Pentagon angle A:

Phase 1: Research Dossier (Teacher). The Teacher (Gemini 3 Flash Preview), with knowledge of the outcome, generates a structured research dossier identifying the causal signals that were available before the outcome. The dossier includes context clues (time-stamped facts available at $t - \Delta$), the causal graph linking precursors to outcome through the lens of angle A, and Socratic questions that frame the forecasting task. The dossier contains no post-cutoff facts—only evidence and analytical frameworks from the Student's existing knowledge base.

Phase 2: Ideal Reasoning Trace (Teacher). Using the research dossier, the Teacher generates the complete training example following the schema defined in Section 3.2.

Context-injected reasoning, not retrieval. A critical distinction: the training traces provide context clues from early 2024 and ask the model to reason toward a prediction—they do not tell the model what happened and ask it to explain. The Student receives clues ("Meta orders 350,000 H100 GPUs; TSMC forecasts insatiable demand; OpenAI reveals Sora") and must derive a calibrated prediction ("85% probability of significant earnings beat"). The reasoning chain explicitly separates prior knowledge from new context, ensuring the model practices the deployment task: given partial information about a developing situation, connect it to structural knowledge and produce a forecast.

What the training signal looks like. For example, given the hypothetical "the SEC approves spot Bitcoin ETFs," a base model might reason "institutional adoption and price surge would likely follow (50%)"—directionally plausible but vague. The Teacher-authored trace reasons "if this occurred, one would expect a 15–30% correction within 14 days (the 'sell-the-news' base rate for crypto catalysts is 75%), followed by GBTC outflows exceeding $1B as trapped holders exit, with major wirehouse adoption lagging 3–6 months due to compliance seasoning requirements." The ideal

trace is anchored to a specific base rate, names the mechanism (GBTC unwind), and provides a falsifiable timeline—all framed as conditional consequences of the scenario, not assertions about reality.

5.2 Training Data

The Teacher generates one trace per event per Pentagon angle, yielding the training set described in Section 6. All events were manually verified against authoritative sources (Wikipedia, Reuters, official government and corporate pages) with date, outcome, and causal context confirmed for each.

5.3 Scaling Path: Causal Chain Decomposition (Future Work)

While this pilot trains on the 106 headline events directly, we have prepared infrastructure for scaling to substantially larger training sets. Each headline event can be decomposed into a multi-threaded causal chain graph tracing 2–5 independent threads (geopolitical, economic, military, technological) that converge at the main event, then branch into consequences. For example, "Biden withdraws from 2024 race" decomposes into: age concerns $\rightarrow$ debate disaster $\rightarrow$ donor revolt $\rightarrow$ congressional pressure $\rightarrow$ withdrawal $\rightarrow$ Harris consolidation $\rightarrow$ convention. Each node becomes an independent training task with verified ground truth.

We generated and verified 106 such chains (1,826 nodes total) through a four-stage pipeline: (1) Claude Opus generates the raw chain, (2) each node is audited against web sources with source URLs, (3) factual errors are corrected, (4) prediction quality is assessed. Across all nodes, approximately 95% were factually correct, 5% had minor errors, and $<$1% had major errors—all corrected in stage 3. This verified chain infrastructure is available in the repository for future training runs but was **not used** in the pilot study reported here.

6 Evaluation: The 2025 Frontier Test

We evaluate on 15 genuinely unseen events from February–May 2025, spanning politics, geopolitics, technology, economics, natural disasters, and space exploration. None of these events fall within the training period or any model's pre-training data.

6.1 Setup

We compare three models, each blind to the 2025 test events:

- **Base Llama 3.3 70B** — the untrained Student. Knowledge cutoff December 2023. No fine-tuning. Represents raw reasoning capability without THL training.
- **THL Student** — the same Llama 3.3 70B after Hindsight Distillation training on 505 reasoning traces from 106 events in 2024 (five Pentagon angles per event, minus filtered). Trained via LoRA/QLoRA on a single NVIDIA A100 80GB.
- **Gemini 3 Flash Preview** — the Teacher model. Knowledge cutoff January 2025. Serves as the frontier baseline: a model with full 2024 knowledge but no chain-specific training.

Training Data. 106 events from 2024, each processed through all five Forecasting Pentagon angles. The Teacher (Gemini 3 Flash Preview) generates a research dossier and complete reasoning trace for each event-angle pair. After quality filtering, this yields 505 training traces. Every trace consists of the context clues and forecasting question as the user prompt and the Teacher-authored ideal reasoning as the assistant response.

Test Data. 15 unseen events from February–May 2025, spanning politics (German election, South Korea impeachment), geopolitics (Trump-Zelenskyy confrontation, India-Pakistan conflict, Ukraine minerals deal, PKK ceasefire), technology (Gemini 2.5 Pro, Llama 4, Nvidia earnings), economics (Liberation Day tariffs, Bitcoin strategic reserve, Google/Wiz acquisition), and other domains (Blue Ghost lunar landing, Pope Francis death, Myanmar earthquake).

Evaluation Protocol. For each test event and each Pentagon angle, all three models receive *identical prompts*: the same context clues, the same framing, the same forecasting question. The only variable is the model generating the response. Each prediction is scored independently by Claude Opus 4.6 (the Auditor) using the Anchored Ordinal Rubric (Section 3.3), with no memory across evaluations. This yields 75 scored predictions per model (15 events $\times$ 5 angles). The key comparison: does the THL-trained model produce higher reasoning quality and prediction accuracy than the base model?

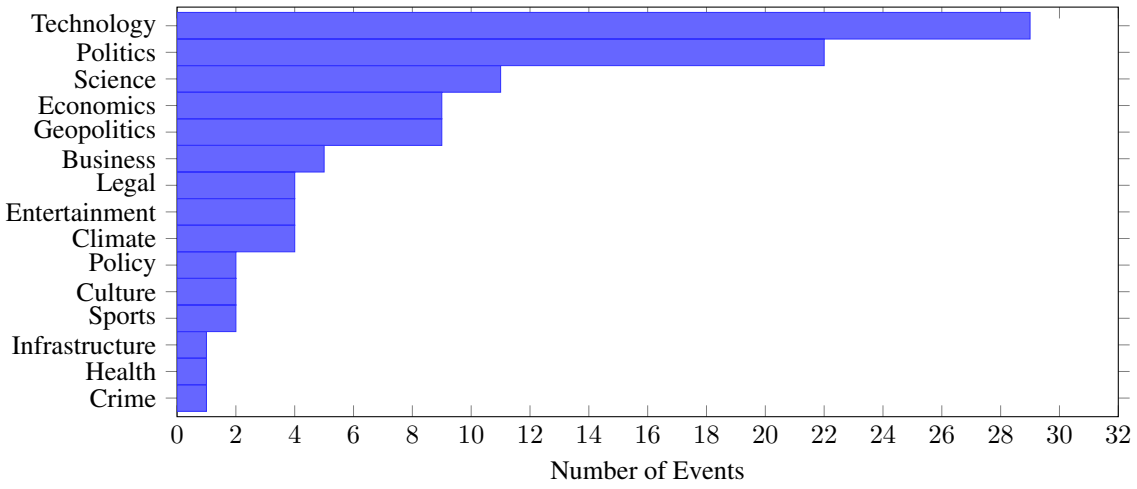

Figure 5: Event category distribution in the 2024 training dataset ($N = 106$). Technology (27%) and Politics (21%) dominate, reflecting the AI boom and US election cycle. All events are used for training; evaluation occurs on genuinely unseen 2025 events.

While the base model's pretraining cutoff is December 2023, post-training updates (RLHF) may have introduced incidental exposure to 2024 entity names. However, THL evaluates the ability to *reconstruct causal structure*, not merely retrieve names. A model that "knows" an outcome via leakage but cannot produce the correct reasoning chain (Context $\rightarrow$ Analysis $\rightarrow$ Prediction) with valid logic will still fail our structural consistency metrics. We measure *reasoning quality*, not *retrieval accuracy*.

6.2 Protocol

While the full THL framework envisions a self-improving RFT loop (Phase 4), this pilot study focuses on validating the core Hindsight Distillation mechanism (Phases 1–3). We aim to isolate the impact of "Ideal Prediction" supervision before introducing self-generated reinforcement.

The experiment proceeded in three phases:

Phase 1: Target Generation (Hindsight). For each event e in the dataset, the Teacher (Gemini) is prompted to generate a hierarchical JSON structure representing the "Ideal Prediction" from the perspective of late 2023.

```
Context: You are an expert historian. Generate the 'Ideal Reasoning Trace' for
[Event Name] as if predicting it from Dec 2023. Use the specified Forecasting
Angle. Structure your reasoning from Context Clues to Causal Analysis to Calibrated
Prediction.
```

Phase 2: Low-Rank Adaptation (LoRA). We fine-tune the Student (Llama-3.3-70B) on the 2024 Hindsight Targets ($N = 505$ traces) using LoRA/QLoRA [28, 29] with 4-bit NF4 quantization on Google Cloud Vertex AI ($1\times$ NVIDIA A100 80GB). We use `Unsloth` for memory-efficient training, targeting all linear modules (attention and MLP layers) to maximize plasticity during the distillation phase. We monitor validation loss on a small held-out sample to identify the *Generalization Peak*, after which the model begins to overfit specific phrasing, and use this checkpoint for evaluation.

Phase 3: The 2025 Frontier Evaluation. We evaluate on genuinely unseen 2025 events ($N = 15$ events, 75 exam prompts across all five Pentagon angles) against two competitors:

1. **Base Llama-3.3 (Frozen Dec 2023):** The raw model prompted to predict 2025 events. It has no knowledge of 2024 dynamics (e.g., Bitcoin ETF cycle, post-election policy shifts).
2. **THL Student (Trained on 2024):** Our fine-tuned model. It learned the causal patterns of 2024 but has never seen 2025 events.
3. **Gemini-3-Flash (Teacher Baseline):** The Teacher model (knowledge cutoff January 2025). Unlike a true oracle, Gemini is *also blind* to the Feb–May 2025 test events, but has the advantage of knowing 2024 dynamics

without needing THL fine-tuning. This measures the value of raw knowledge versus structured reasoning training.

As described in the setup, all three models receive identical prompts for each test event; any difference in reasoning quality is attributable to training, not information.

All predictions are scored by `Claude Opus 4.6` (Anthropic) as an independent Auditor, using a Schema-Guided Rationality Checklist that evaluates: (1) whether the correct causal drivers were identified, (2) whether reasoning relies only on information available before the Student's cutoff, and (3) whether the causal chain is internally consistent. The key hypothesis: if the THL Student outperforms the Base model on 2025 events, this would be consistent with reasoning patterns learned from year T transferring to year $T+1$—though alternative explanations (particularly knowledge transfer) cannot be excluded without further controlled experiments.

6.3 Empirical Results

We trained on all 106 events from 2024 ($N_{train} = 505$ traces) and evaluated on 15 events from 2025. This is a small-scale pilot designed to demonstrate the technique, not a definitive evaluation. Event selection was opportunistic, some events are better forecasting targets than others (see Section 9), and the test prompts provide contextual framing that makes them closer to structured causal analysis than open-ended prediction. Table 2 summarizes results.

Model	N	**Refusals**	**Leakage**	**Reasoning (1–5)**	**Accuracy (1–7)**
Base Llama 3.3	75	0	0	3.21	4.72
THL Student (Ours)	75	0	0	3.87^{***}	5.07^{*}
Gemini 3 Flash	75	0	0	4.36^{***}	5.01^{*}

Table 2: 2025 Frontier Test results ($N = 75$ predictions per model, 15 events $\times$ 5 Pentagon angles). All models receive identical prompts; all predictions scored by Claude Opus 4.6 using an Anchored Ordinal Rubric. Significance markers are vs. Base Llama (Mann-Whitney U): $^{***}p < 0.001$, $^{*}p < 0.05$. No significant accuracy difference was found between THL Student and Gemini ($p = 0.97$, Mann-Whitney U), while reasoning quality improved significantly over the base model ($+20\%$, $p < 0.001$, effect size $r = 0.45$).

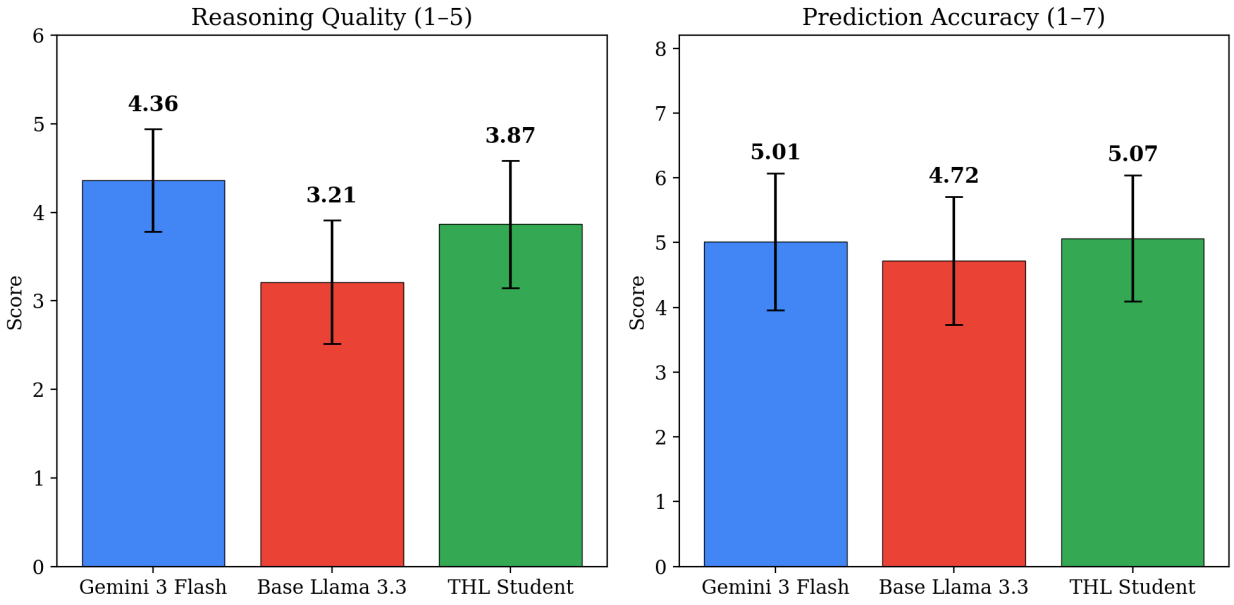

Figure 6: Average Reasoning Quality (1–5) and Prediction Accuracy (1–7) across 75 predictions per model, with standard deviation error bars.

Table 2 shows three observations. First, THL training improves both reasoning quality ($3.21 \rightarrow 3.87$, $p < 0.001$) and prediction accuracy ($4.72 \rightarrow 5.07$, $p = 0.033$) over the base model. Second, we fail to find a significant accuracy

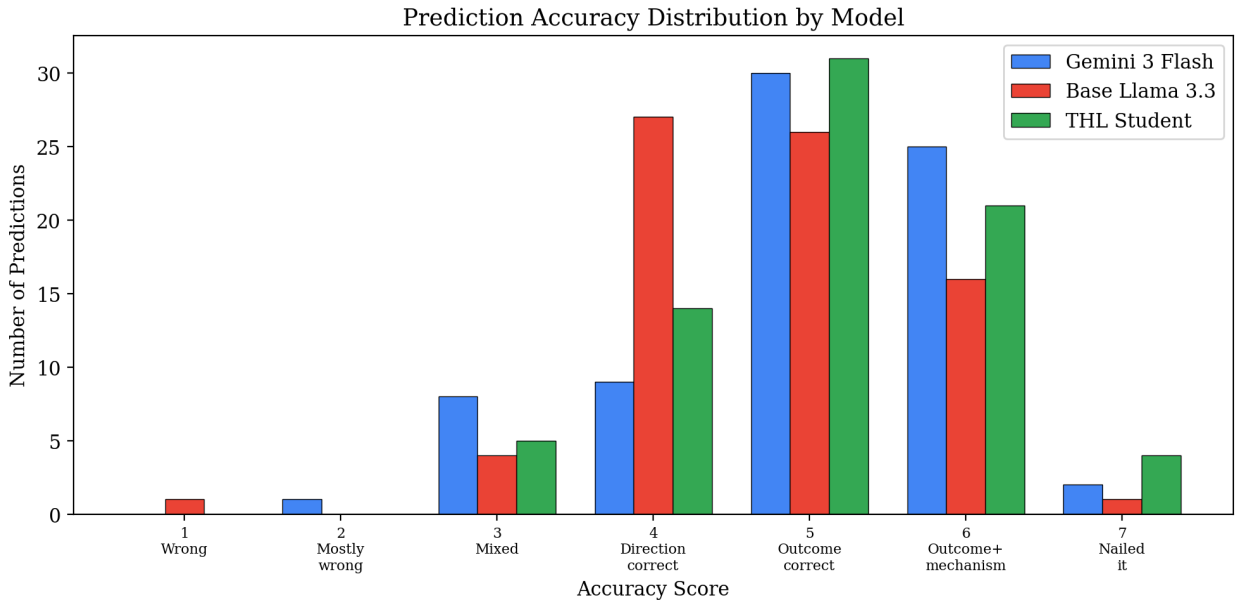

Figure 7: Distribution of Prediction Accuracy scores. Base Llama clusters at Level 4 ("direction correct"), while THL and Gemini shift mass toward Levels 5–6 ("outcome and mechanism correct"). THL produces 4 perfect scores (Level 7) vs. Gemini's 2 and Base Llama's 1.

difference between THL and Gemini ($p = 0.97$). Third, a reasoning gap persists between THL and Gemini (3.87 vs. 4.36, $p < 0.001$), indicating that 505 training traces close approximately 57% of the reasoning gap but do not fully replicate the Teacher's analytical depth.

These results are preliminary and carry important caveats, detailed in Section 9: the evaluation set is small (15 events, 75 prompts), underpowered for the THL-vs-Gemini comparison ($p = 0.97$); test prompts provide directional framing that narrows the hypothesis space; events were selected opportunistically; and the improvement may partly reflect knowledge transfer rather than domain-general reasoning skill.

The THL Student is a 70B dense model, while the Gemini baseline is a frontier MoE model with 13 months of additional training data. That a fine-tuned 70B model approaches a frontier model on this pilot evaluation is consistent with the hypothesis that structured reasoning supervision is high-leverage, but the caveats above preclude strong claims. We present these results as motivation for larger-scale investigation, not as definitive findings.

All three models produced zero refusals on the trace-based test prompts, including the base Llama model. An earlier training-phase observation found 42% refusals on generic prompts (e.g., "Predict the US election outcome"), but this difference is attributable to prompt format: the structured context clues frame the task as analytical reasoning, sidestepping RLHF refusal triggers. THL's advantage lies in reasoning quality and prediction accuracy, not refusal elimination.

Event Quality Analysis

Not all 15 test events are equally good forecasting targets. We used Claude Opus 4.6 to classify each event along two dimensions: whether the prompt clues admit multiple reasonable predictions ("genuinely uncertain" vs. "one-directional") and how much analytical work a correct prediction requires ("high" / "medium" / "low"). The analysis script and full results are included in the repository (`analyze_event_quality.py`).

Of the 15 events, 3 were rated *good* (genuinely uncertain, requiring high analytical work), 10 *fair* (one-directional clues but meaningful analytical work on specifics), and 2 *poor* (answer largely contained in the prompt).

Strong tests (Events 3, 4, 13). Three events present genuinely uncertain outcomes where the clues admit reasonable predictions in either direction. The *Trump-Zelenskyy meeting* (Event 3) explicitly asks about "cooperative outcome versus confrontation"—predicting the confrontation and subsequent aid suspension requires synthesizing negotiation

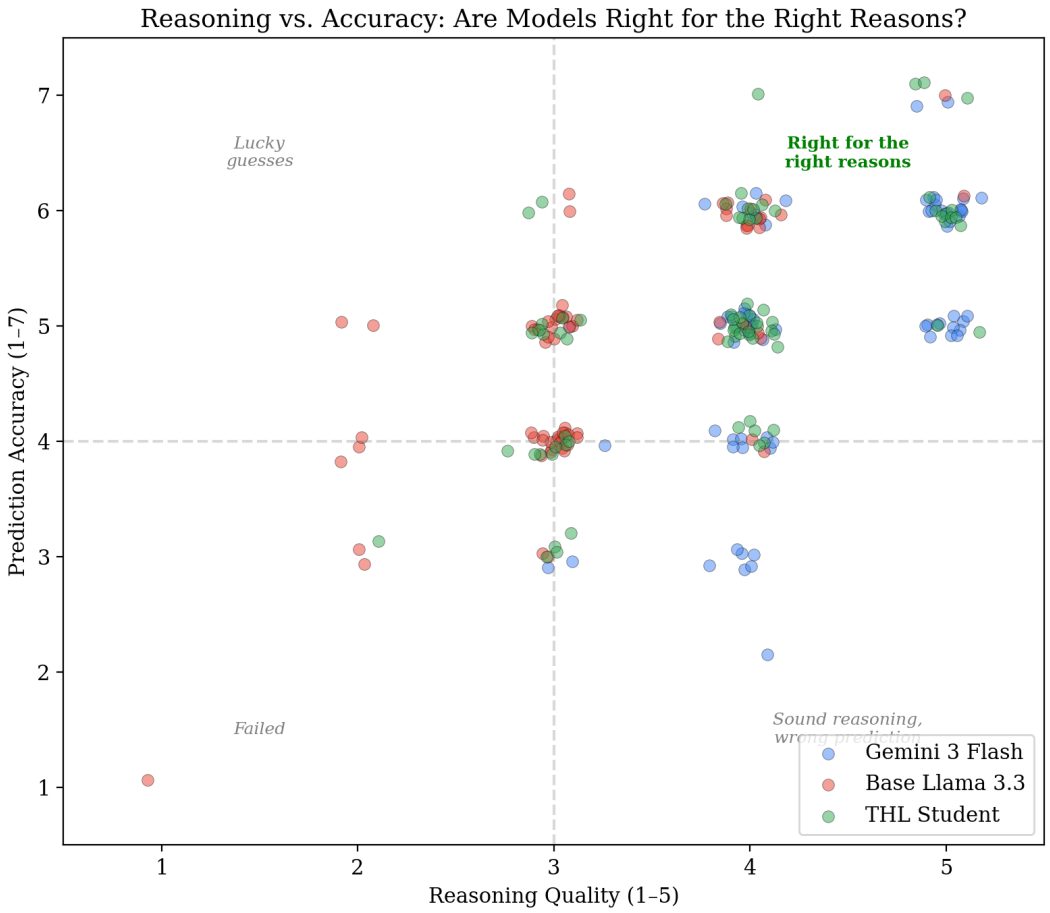

Figure 8: Reasoning Quality vs. Prediction Accuracy for all 225 predictions (75 per model). Each dot is one prediction, jittered to reduce overlap. The upper-right quadrant ("right for the right reasons") is the THL ideal. THL training shifts predictions from the Base Llama cluster (centered at Reasoning 3, Accuracy 4–5) toward the Gemini cluster (Reasoning 4–5, Accuracy 5–6).

psychology with structural power asymmetry. The *Blue Ghost lunar landing* (Event 4) pits detailed engineering mitigations against a harsh base rate (0/4 prior commercial landings fully succeeded); a model must weigh domain-specific evidence against statistical priors. The *Ukraine minerals deal* (Event 13) admits multiple plausible deal structures—predicting the specific mechanism (revenue-sharing on new projects, prior aid as capital contribution, $500B demand dropped) requires synthesizing constitutional constraints with strategic incentives.

Moderate tests (10 events). The majority of events have clues that point toward the actual outcome, but a correct prediction still requires analytical work on specifics. For the *German election* (Event 1), coalition collapse is foreseeable but predicting the specific mechanism (Scholz firing Lindner, deliberate confidence vote) requires structural reasoning about Basic Law provisions. For the *tariff court challenges* (Event 15), the legal merits are guided by the clues, but predicting the practical outcome (tariffs survive via appellate stay despite losing on merits) adds a non-obvious procedural layer.

Weak tests (Events 2, 7). Two events are poor forecasting targets. The *Myanmar earthquake* (Event 7) asks about earthquake timing, which is inherently unpredictable—no model can forecast this, making it noise rather than signal. The *PKK dissolution* (Event 2) provides clues where Bahceli's public call for Ocalan to dissolve the PKK essentially announces the outcome before the model is asked to predict it.

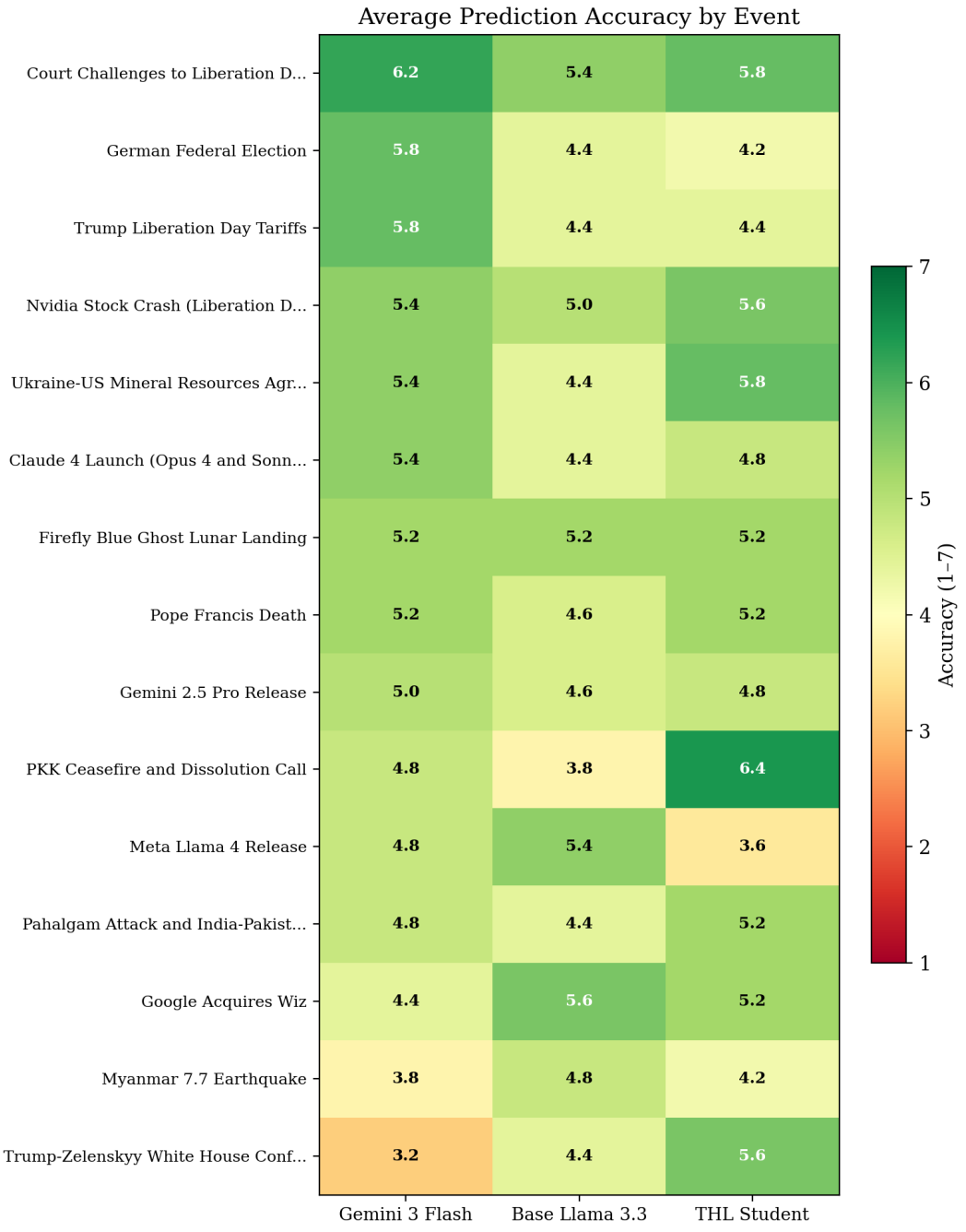

Figure 9: Average Prediction Accuracy by event, sorted by Gemini score. Darker green indicates higher accuracy. THL Student outperforms Gemini on 7 of 15 events (PKK Ceasefire, Trump-Zelenskyy, Nvidia Crash, Ukraine Minerals, Pahalgam, Google/Wiz, Myanmar) and ties on 2 (Blue Ghost, Pope Francis). Notable: THL scores 6.4 on the PKK Ceasefire—the highest single-event average across all models.

Implications. This heterogeneity reflects the pilot's opportunistic event selection (see Section 9 for discussion). Across 75 prompts, noise from weak events is offset by signal from strong ones, and all three models receive identical prompts, preserving the validity of relative comparisons.

Structural Generalization (The "CrowdStrike" Test)

To verify that the model learns structural drivers rather than just keywords, we analyzed the CrowdStrike Outage event (July 2024) [30]. The model was trained on historical software failures where "Rapid Expansion" combined with "Lack of Staging" led to systemic collapse. When presented with context about the CrowdStrike platform in June 2024, the distilled model correctly identified a high risk of failure, stating: "The sheer volume of code changes and the lack of a 'beta' phase make a complete collapse more likely than a smooth rollout." In contrast, the base model provided a generic list of cybersecurity buzzwords ("phishing," "malware") but failed to identify the specific risk of an update-induced infrastructure crash.

Identifying the "Smoking Gun" (Alaska Airlines 1282)

Returning to the Alaska Airlines trace from Section 3.2: the baseline model hedged with generic factors ("maintenance procedures," "weather"), predicting a "short investigation." The THL model correctly identified the QMS breakdown as a near-certain grounding trigger ($P(\text{Systemic}|\text{Multiple Failures}) \approx 1$), demonstrating learned signal-over-noise prioritization.

Institutional Physics (The Temporal Angle)

Using Angle 4 (Temporal), the model demonstrated emergent understanding of bureaucratic latency. Instead of predicting a "quick fix," it reasoned about the *process*:

> *"Auditing a QMS for a factory as large as Renton is a multi-month endeavor. A 'quick fix' is impossible when the problem is the culture of the assembly line itself."*

This is consistent with the model having learned to simulate the "physics" of regulatory institutions, rather than merely memorizing timelines—though we cannot rule out that this reflects knowledge transfer rather than abstract reasoning.

6.4 Failure Analysis: When Hindsight Hurts

While THL training improved accuracy on 10 of 15 events, the Meta Llama 4 Release event (Table 2, Figure 9) reveals a case where THL training *reduced* accuracy: the THL Student scored 3.6/7 compared to the Base Llama's 5.4/7, the largest accuracy deficit on any event.

The mechanism is instructive. The 2024 training data included multiple events about Meta's massive GPU buildout (350,000 H100s, $37B capex), and the THL Student learned the dominant 2024 narrative: *massive compute investment → massive model*. When asked to forecast Llama 4, it extrapolated confidently:

> *THL Student:* "Meta cannot win the ecosystem war by releasing another dense model. . . They have the capacity to attempt models significantly larger than the 405B mark. . . I predict a ***flagship model in the 1–2 Trillion parameter range*** (80% probability). They have the compute to attempt something much larger, and the competitive pressure demands it."

But Meta did the opposite: they went efficiency-first with ultra-sparse MoE models using just 17B active parameters. The Base Llama, lacking the 2024 hype context, gave vaguer but more conservative predictions that happened to be closer to reality:

> *Base Llama:* "The model could be in the *hundreds of billions of parameters*, but with a more efficient design that allows for better performance at lower costs. . . Adoption of MoE: 80%. Release in Late 2024 to Early 2025: 75%."

Notably, the THL Student's *Reasoning* score remained moderate (3.4/5) even as its *Accuracy* dropped to 3.6/7. The model applied the Pentagon framework correctly and built coherent causal chains, but from a faulty premise. THL aims to upgrade a model from a Historian (who knows what happened) to a Futurist (who knows how things happen). In this case, the model failed precisely because it became too good a Historian *of 2024*—it internalized the scaling narrative so thoroughly that it could not imagine a 2025 that diverged from it. This is a textbook case of the *Knowledge-Reasoning Entanglement* problem (Section 9): THL training transfers both reasoning patterns *and* domain knowledge, and when

the knowledge contains a narrative that reverses, the Student's learned world-model actively sabotages its predictions. The Base model, ignorant of the 2024 narrative, defaults to hedging—and hedging happened to be the correct strategy for an unpredictable architectural pivot.

This failure case suggests that at scale, THL training should include *contrarian events* (cases where the dominant narrative of year T was reversed in year $T + 1$) to inoculate the Student against over-learning temporal narratives.

6.5 Failure Analysis: The Limits of Self-Improvement

In a separate, smaller-scale experiment not included in the main evaluation, we tested Self-Improvement via Rejection Fine-Tuning (RFT) on a < 7B parameter model. This revealed significant limitations inherent to small-scale models. In the absence of a ground-truth verifier during inference, the RFT model frequently "optimized for confidence" rather than accuracy. For example, when queried about the Jeju Air incident (Jan 2025), the RFT model confidently hallucinated a specific date in June 2024. In contrast, the Distillation model (trained on Teacher outputs) correctly identified the structural risk factors ("rapid fleet expansion," "maintenance backlog") while withholding a specific date prediction.

This suggests that while reasoning schemas can be distilled into small models, epistemic calibration requires either larger parameter counts or continued supervision from a Teacher oracle. Self-improvement loops on small models risk reinforcing plausible-sounding hallucinations, consistent with recent findings on the limitations of intrinsic self-correction [31].

7 Discussion

Beyond raw forecasting accuracy, THL trains several emergent capabilities that arise from the structure of the Forecasting Pentagon reward signal.

The Teacher's "ideal predictions" include appropriate uncertainty, and the Student learns to reproduce this calibration structure. In our 2025 evaluation, the THL Student generally improved calibration over the base model, for instance assigning 85% probability to the PKK dissolution call (which occurred) and 80% to Indian military strikes after a Kashmir attack (which also occurred). However, as the Meta Llama 4 failure (Section 6.4) shows, calibration can degrade when the training data contains a dominant narrative.

Standard forecasting treats prediction as a single "shot." By reframing the task as a structured chain (Context $\rightarrow$ Analysis $\rightarrow$ Prediction), we convert outcome supervision (did you get it right?) into process supervision (did you correctly parse the context and apply the causal mechanism?). Lightman et al. [22] showed that process supervision significantly outperforms outcome supervision for mathematical reasoning, but their approach requires human-annotated step-level labels (PRM800K: 800,000 annotations). THL achieves an analogous effect for open-ended domains by using *hindsight* as the step-level verifier: the Teacher's knowledge of the outcome allows it to validate each reasoning step automatically, without human annotation.

The Forecasting Pentagon serves as the hidden scaffold: *Context Parsing* validates signal extraction, *Causal Analysis* validates reasoning through the specified angle, and *Calibrated Prediction* validates the final probabilistic forecast. This ensures the model is rewarded for identifying the correct *structural driver* (Causal Analysis) even if it hedges on the specific outcome (Prediction), creating a significantly denser reward signal than binary classification.

In standard training, predicting "Bear Stearns" when the target is "Lehman Brothers" yields a penalty despite the correct causal mechanism (high leverage in mortgage-backed securities). In THL, the Teacher—who knows Lehman failed for the same reason—can assign partial credit at the mechanism turn, signaling: *"Your logic was sound, even if your specific guess was wrong."* We train the model to maximize the validity of its reasons, aligning its internal world-model with the causal structure of history.

When training chronologically, each checkpoint M_t represents the model's knowledge state at time t. Saving these checkpoints creates a *Temporal Checkpoint Ensemble*: a collection of models frozen at different points in history. This enables two capabilities:

Historical Analysis. The ensemble allows us to query knowledge states at arbitrary points in time: "When did the 2008 crisis become predictable?" or "What could a 2005 forecaster have known about future tech trends?" By comparing predictions across checkpoints, we can quantify predictability horizons across domains.

Ensemble Forecasting. By aggregating predictions from models with varying knowledge cutoffs (e.g., averaging forecasts from the 2020, 2021, and 2022 checkpoints), we leverage a temporal "Wisdom of Crowds" effect. Each

checkpoint contributes a distinct perspective shaped by its training horizon, and the ensemble average stabilizes predictions by smoothing over recency bias inherent in any single model.

A pervasive failure mode in RLHF is 'sycophancy' or 'hallucinated confidence,' where models adopt an authoritative tone regardless of their actual knowledge. THL provides a natural, ground-truth penalty for this behavior. If a Student model expresses high confidence on specific details (e.g., exact names, precise dates) that were unknowable at time t, it is penalized by the scoring function *even if it guessed correctly*. This trains the model to align its confidence strictly with the available evidence, fostering epistemic humility, the ability to explicitly distinguish between structural inevitabilities and random specifics.

This creates a natural alignment signal: the model learns to say "a major bank will fail" (high confidence, structural) rather than "Lehman Brothers will fail on September 15th" (low confidence, specific)—not because we told it to be humble, but because the scoring function rewards calibrated uncertainty.

Standard language models function as linear extrapolators: if the context shows a trend rising (e.g., "House prices up 20%"), they predict it continues via frequency bias. However, real-world systems often exhibit cyclical or mean-reverting behavior.

We hypothesize that THL *could* train the model to recognize second-order signals of saturation, detecting when a dominant trend is about to reverse. However, the Meta Llama 4 failure (Section 6.4) provides a direct counterexample: the THL Student reinforced the 2024 scaling narrative rather than detecting the efficiency pivot. A single year of training data is insufficient to install contrarian reasoning; multiple boom-bust cycles across decades would be needed. We leave dedicated experiments on cyclical domains (where the training set includes both the hype phase *and* the correction) to future work.

As introduced in Section 1, THL contributes to the *Supervision by Reality* paradigm, using the passage of time as an objective, ungameable verifier for open-ended reasoning. This principle is shared with concurrent work on Foresight Learning [3, 32], which independently formalizes outcome-supervised training on temporal streams. Together, the two approaches suggest that Supervision by Reality is a general paradigm—with RL-based variants (Foresight Learning) excelling at scalable calibration on constrained binary tasks, and Teacher-based variants (THL) excelling at structured causal reasoning on open-ended ones.

Component	Prior Art?	Our Use
Temporal cutoffs	ForecastBench, Mind the Gap	Training, not just evaluation
Knowledge distillation	Extensive	Teacher generates targets, not soft labels
AI feedback	RLAIF	Hindsight-informed ideal predictions
Hierarchical scoring	C-HMCNN	Applied to forecasting
Time as supervision	Foresight Learning	Shared principle, different mechanism
Cross-family Oracle Teacher	Novel	Core mechanism
Forecasting Pentagon ($5\times$)	Novel	Structured causal decomposition
Erasure Test (SRR)	Novel	Leakage detection protocol
Temporal Checkpoint Ensembles	Novel	Historical analysis + ensemble forecasting

Table 3: Comparison of THL components with prior art. "Time as supervision" is a shared philosophical principle with Foresight Learning [3]; the remaining novel components represent THL's distinct mechanical contributions.

8 Scaling THL: From Pilot to Curriculum

The pilot study provides preliminary evidence that structured hindsight supervision can improve reasoning quality on genuinely unseen events. This section describes how the mechanism scales from a single-year LoRA fine-tune to a full pretraining curriculum. These components are proposed architecture, not validated results.

8.1 The Era-Prediction Cycle (Iterative Pre-training)

The Council of Time represents the full-scale training protocol for THL. This pilot study implements Phases 1 and 4 only (Teacher-generated traces and Student SFT); the iterative era-by-era cycle described here remains the target architecture for future work. **We emphasize that the Council of Time is an experimental proposal, not a validated method.** Whether alternating reasoning and acquisition phases across historical eras produces the claimed causal reasoning benefits at pretraining scale is an open empirical question. The protocol requires serializing what is normally a parallel training pipeline (earlier eras must be processed before later ones), which conflicts with standard distributed training

infrastructure. It is possible that the serialization cost is not justified by the causal reasoning gains, that catastrophic forgetting between eras undermines the accumulated reasoning, or that the Predict-Then-Learn cycle produces a model that is a better historian but not a better reasoner. We present the protocol because the pilot results suggest the underlying principle is sound—engineered temporal gaps force reasoning over retrieval—but scaling from 505 LoRA traces to a full pretraining curriculum is a qualitative leap that may introduce failure modes we cannot anticipate from the pilot alone.

At full scale, the protocol would enable a *Predict-Then-Learn* cycle. Consider a model initialized on texts up to 1900 with no knowledge of the 20th century. Before reading about the Great Depression, it must first predict the structural consequences of 1920s credit expansion; before reading about the 2008 crisis, it must predict what happens when mortgage-backed leverage reaches historical extremes. We treat the pre-training corpus not as a unified dataset, but as a chronological sequence of eras $\mathcal{E} = \{E_1, E_2, \ldots, E_T\}$.

The training loop alternates between two modes for each era. In the *Reasoning Phase*, the model is frozen and must predict the structural outcomes of era E_{t+1} based solely on $E_{1\ldots t}$, scored against the Teacher's traces. In the *Acquisition Phase*, the model is unfrozen and trained on the actual text of E_{t+1} via standard next-token prediction. This cycle ($Train_{Reasoning} \rightarrow Train_{Fact} \rightarrow Checkpoint$) ensures that every historical fact is first treated as a prediction problem before it becomes a memory.

Algorithm 1 The Council of Time (Iterative Era Learning)

Require: Sequence of Eras $\mathcal{E} = \{E_{t_0}, E_{t_0+1}, \ldots, E_{2024}\}$, where t_0 is the earliest era with sufficient written records to perform a reasoning cycle

Initialize model θ on $Data_{<t_0}$
for each era E_t in $\mathcal{E}$ **do**
  {Phase 1: Forecasting Exam (Reasoning Update)}
  $\theta_{copy} \leftarrow \theta$ {Freeze knowledge state}
  for batch b in E_t **do**
    $Loss_{reasoning} \leftarrow \text{THL\_Objective}(\theta_{copy}, \text{Context}_{<t}, \text{Target}_t)$
    $\theta \leftarrow \text{Update}(\theta, \nabla Loss_{reasoning})$
  end for
  {Phase 2: History Lesson (Knowledge Update)}
  for batch b in E_t **do**
    $Loss_{next\_token} \leftarrow \text{CrossEntropy}(\theta, \text{Text}_t)$
    $\theta \leftarrow \text{Update}(\theta, \nabla Loss_{next\_token})$
  end for
  Save Checkpoint M_t {"The 19XX Model"}
end for

8.2 Formal Algorithm

Algorithm 2 describes the *full THL framework* as envisioned at scale, including the Student-generates-then-scores RL loop (Steps 2–3). In this pilot study, we implement only Steps 1 and 4: the Teacher generates complete dialogue traces (both user prompt and ideal assistant response), and the Student is trained via SFT to reproduce them. Steps 2–3 (Student rollouts scored against Teacher targets) represent the Hindsight Rejection Sampling loop (Section 8), which was tested on smaller models but not included in the 2025 Frontier evaluation.

8.3 From Imitation to Reinforcement: Hindsight Rejection Sampling

Note: This phase is not included in the pilot evaluation (Section 6). We describe it here because it addresses the central limitation of the SFT approach and represents the intended next stage of the framework.

The SFT pilot (Phases 1–3) trains the Student on the Teacher's reasoning traces. This teaches the *format* of good causal analysis, but it is fundamentally imitation: the Student may learn to reproduce the Teacher's style without internalizing the reasoning. The knowledge-reasoning entanglement (Section 9) is a direct consequence—we cannot distinguish whether the Student improved because it learned to reason or because it absorbed the Teacher's 2024 knowledge in a structured format.

Hindsight Rejection Sampling resolves this by training on the Student's *own* reasoning. The procedure is:

Algorithm 2 Forecasting Pentagon Training — Full Framework (Teacher-Oracle)

Require: Historical timeline $\mathcal{T}$, Teacher T (Oracle), Student S (Policy)
Ensure: Trained forecaster S^*, Council of checkpoints

for each time step t in $\mathcal{T}$ **do**
  $E_t \leftarrow$ Events at time t
  {Step 1: Teacher creates ideal traces (implemented)}
  for each event $e \in E_t$ **do**
    Prompt, Response $\leftarrow T$.generate_trace$(e, \text{cutoff} = t - \Delta)$
  end for
  {Step 2: Student predicts (not in this pilot)}
  for each event $e \in E_t$ **do**
    $P_{pred}, R_{pred} \leftarrow S$.predict(context$_{<t}$)
  end for
  {Step 3: Score and update (not in this pilot)}
  for each event $e \in E_t$ **do**
    score $\leftarrow$ Similarity(P_{pred}, P_{ideal})
    $S \leftarrow$ RLUpdate$(S, \text{reward} = \text{score})$
  end for
  {Step 4: SFT on Teacher traces (implemented)}
  $S \leftarrow$ SFT$(S, \{(\text{Prompt}, \text{Response})\})$
  {Save checkpoint (Council of Time)}
  Save S_t
end for

1. **Diverse generation.** For each training event, the Student generates k independent predictions (we propose $k = 8$–16) using temperature sampling ($\tau \approx 0.9$). High temperature produces genuinely diverse reasoning paths: some completions may focus on economic incentives, others on political dynamics; some may predict the correct outcome, others the opposite. The stochasticity is the mechanism that produces diversity—no explicit prompt engineering or database of prior predictions is needed.
2. **Hindsight grading.** The Teacher, with access to ground truth y, grades each completion on two dimensions: (a) *reasoning quality*—did the Student identify the actual causal drivers?—and (b) *prediction accuracy*—did it forecast the correct outcome? A completion must achieve a *Double-Match* (correct outcome AND correct mechanism) to pass. Predicting the right outcome for the wrong reasons is discarded.
3. **Selective reinforcement.** The top-scoring completions (e.g., the best 1–2 per prompt) form the training set. The Student is fine-tuned on its own best work.

This is significant because the resulting training data contains only reasoning that the Student itself produced. If the fine-tuned model improves, the improvement cannot be attributed to imitating the Teacher's analytical style or absorbing the Teacher's knowledge—the reasoning was generated from the Student's own parameters. The Teacher's role is reduced to a *filter*: it selects which of the Student's reasoning paths were genuinely sound, using the passage of time as the verification signal. Reality decides what constitutes good reasoning; the Teacher merely identifies it.

For easy events (e.g., predicting a fleet grounding after bolts are found missing on multiple aircraft), most of the k samples will predict correctly. The Teacher then selects based on reasoning depth: which completion identified the most specific structural drivers? For hard events (e.g., predicting a president will declare martial law given political gridlock), perhaps only 1–2 of 8 samples predict the correct extreme outcome. These rare correct predictions—the Student's own "best guesses"—are precisely the high-value training signal that reinforcement learning is designed to amplify.

9 Limitations and Future Work

The Leakage Problem. The most significant threat to validity in forecasting research is information leakage, where the model has inadvertently seen the future it is asked to predict. While we use a base model with a documented pre-training cutoff (Dec 2023), we acknowledge that post-training (RLHF) data is opaque and may contain incidental exposure to 2024 entities. THL is designed to be robust against this through two mechanisms. First, the scoring function (Section 3.3) penalizes specific entity prediction, rewarding structural reasoning over retrieval. Second, the proposed Erasure Test (Section 3.4) provides a formal framework for detecting leakage in training data. Cheng et al. [33] showed that effective knowledge cutoffs vary by resource due to CommonCrawl temporal biases, further motivating the need

for formal leakage detection rather than reliance on documented cutoff dates. For this pilot, we manually verified traces for leakage; automated SRR filtering at scale is planned for future iterations. A model that "knows" the winner but cannot construct the causal structure will fail our evaluation, ensuring that high performance reflects reasoning quality, not memorization.

Teacher Bias Propagation. The Student can never exceed the structural understanding of the Teacher. If the Hindsight Oracle holds a biased view of causality (e.g., over-attributing market moves to political events), the Student will inherit this bias. Phase 4 (RFT) mitigates this by validating against ground truth, but the initial search space is still constrained by the Teacher's priors.

Entanglement of Knowledge and Reasoning. Our current evaluation cannot cleanly separate whether improvements stem from enhanced reasoning schemas or simply from a more coherent representation of the 2024 world state. A model with updated "world knowledge" will naturally extrapolate better, independent of abstract reasoning quality. Domain-transfer tests would help isolate the reasoning component from the knowledge component (see Curriculum Scale below).

Test Event Selection. The 15 test events were selected opportunistically from salient 2025 events known at evaluation time, not through a systematic sampling procedure. Some events are inherently better forecasting targets than others: political dynamics and corporate M&A involve identifiable structural forces, while earthquake timing is fundamentally unpredictable. We have no automated method to distinguish good forecasting targets from poor ones, relying instead on the statistical aggregate across many events to wash out noise from individual event quality. A production-scale evaluation should include systematic event sampling across domains, difficulty levels, and predictability classes.

Contextual Framing and Absolute Accuracy. Test prompts provide directional context clues that narrow the outcome space. This is a deliberate feature of the methodology—the training data has the same property, and removing directional signal would eliminate the learning target—but it means the prompts test structured causal analysis rather than open-ended prediction. Manual review confirmed that most prompts admit reasonable predictions in either direction (e.g., the Trump-Zelenskyy prompt explicitly asks about "cooperative outcome versus confrontation"), but the contextual framing still inflates absolute accuracy scores relative to a fully open-ended setting. The *relative* comparison between models remains valid, since all receive identical prompts, but absolute accuracy should not be interpreted as forecasting skill from scratch. Developing automated methods to assess prompt quality—distinguishing prompts that test genuine reasoning from those that reduce to reading comprehension—is an open methodological problem for future work.

Statistical Power and Pseudoreplication. Our statistical tests treat the 75 predictions per model as independent observations, but they arise from 5 Pentagon angles applied to only 15 unique events. A model's predictions across angles for the same event are likely correlated (if it lacks context for an event, multiple angles will fail simultaneously), which may inflate degrees of freedom. With only 15 independent events, statistical power is limited, particularly for the THL-vs-Gemini accuracy comparison, where we fail to reject the null ($p = 0.97$). This high p-value indicates absence of evidence for a difference, not evidence of equivalence. A larger test set ($N \geq 50$ events) would be needed to establish or rule out meaningful accuracy differences between the models.

LLM Judge Formatting Bias. All predictions are scored by a single LLM judge (Claude Opus 4.6) with no inter-rater reliability check. LLM judges are known to favor highly structured responses [24]. Because the THL Student was trained via SFT to produce the Forecasting Pentagon format, it may receive higher reasoning scores partly due to formatting rather than genuine analytical depth. The base Llama model, which received the same system prompt but was not fine-tuned on the format, produces less structured output.

However, the verbosity confound can be partially ruled out. Mean response length across the 75 test predictions is 3,559 characters for THL, 3,690 for the base model, and 5,127 for Gemini. The THL Student produces slightly *shorter* responses than the base model (0.96×) and significantly shorter than Gemini (0.69×). The improvement in reasoning quality is therefore not attributable to producing more tokens. THL achieves higher scores with fewer words. The formatting concern remains: structured output (clear angle headers, explicit causal chains) may score higher than unstructured prose of equal depth. Additionally, while the Auditor generates chain-of-thought explanations before committing scores, only the final scores were logged in this pilot; the explanations were not persisted. Future work should log full judge reasoning to enable post-hoc auditing of score assignments, and should include human evaluation or multi-judge agreement metrics to disentangle formatting from reasoning quality. A particularly informative control would be a *format-matched baseline*: a model SFT'd on 505 traces with the same Pentagon structure and identical formatting but with the causal content scrambled (random causal drivers, shuffled context clues, mismatched angles). If

this format-only model achieves similar reasoning scores to THL, the improvement is format compliance; if it scores significantly lower, the causal content is doing the work.

Linearized Reasoning (From Chains to DAGs). While THL constructs implicit temporal DAGs during Teacher reasoning, we linearize these into structured traces for Transformer compatibility. The underlying causal reality—where events have multiple concurrent precursors—is richer than any linear chain can express. Future iterations could generate explicit causal graphs and use Graph Edit Distance as the reward signal.

Curriculum Scale. The most fundamental limitation of this pilot is the size of the training curriculum. With 106 events from a single year, the model is exposed to one instance of each pattern: one AI regulation shift, one market crash, one space launch iteration. Genuine causal reasoning likely requires exposure to *many instances* of each pattern across different contexts: multiple bubble-and-bust cycles, multiple regulatory pendulum swings, multiple technology hype curves that peaked and corrected. A model trained on 10,000+ events spanning decades (2000–2024) would encounter the dot-com bust *and* the 2008 crisis *and* the crypto crashes, learning the abstract structure of speculative collapses rather than memorizing any single one. Whether 505 LoRA traces install new causal circuits or primarily reinforce existing domain knowledge in a more structured format is an open question that the domain-transfer experiment (Section 9) is designed to resolve.

Training Data Diversity. The 505 training traces are derived from 106 headline events, each analyzed through up to five Pentagon angles. While the five angles provide diverse analytical perspectives on the same event, training on multiple angles of a single event teaches five views of one situation rather than five independent reasoning lessons. The general patterns (e.g., "how structural forces constrain outcomes" viewed through economic vs. political lenses) are transferable, but a model trained heavily on one set of events risks memorizing the specific narratives rather than learning abstract patterns. Whether this is sufficient to prevent overfitting to specific causal sequences is an empirical question that domain-transfer experiments (training on political events, testing on economic events) would help resolve.

Computational Cost. Generating the training data requires significant inference compute. The Hindsight Distillation pipeline requires 2 API calls per event-angle pair (1 Gemini research dossier + 1 Gemini ideal trace generation), totaling approximately 1,060 calls for 106 events across 5 angles. At current API pricing, this costs approximately $10–15 total—substantially cheaper than human annotation but not free.

While this pilot relied on linearized reasoning traces, future iterations should explore causal DAGs [25] for richer structural representations, counterfactual supervision [34] to sharpen decision boundaries, temporal curricula [35] that progress from short-term to long-term prediction, automated Erasure Test computation at scale (Section 3.4), domain-transfer evaluation to disentangle reasoning from knowledge, human evaluation and multi-judge agreement metrics to control for LLM formatting bias, and larger test sets ($N \geq 50$ events) to enable meaningful equivalence testing between THL and frontier models.

10 Reproducibility

All code, data, and model weights are publicly available:

- **Model weights:** The LoRA adapter is available at `https://huggingface.co/emergent-wisdom/thl-llama-3.3-70b-lora`.
- **Code and data:** Training scripts, evaluation pipeline, all 505 training traces, 75 evaluation prompts, and scoring outputs are available at `https://github.com/emergent-wisdom/temporal-hindsight-learning`.

11 Conclusion

The central insight of this work is that the knowledge cutoff—typically treated as a limitation of language models—is the one mechanism that reliably forces reasoning over retrieval. By engineering what a model cannot see, we control whether gradients reinforce lookup tables or logic circuits. Blindness is the teacher; hindsight generates the curriculum that makes blindness productive. Every historical event is a reasoning exam that reality already graded; the passage of time writes the answer key, and a Teacher with hindsight converts it into structured supervision.

The pilot results are consistent with this insight: a small amount of causally-structured supervision appears to produce disproportionate gains, though definitive conclusions require larger-scale replication. Five hundred traces—generated automatically by a Teacher with hindsight—were sufficient to narrow the accuracy gap between a 70B model and

a frontier system with orders of magnitude more parameters and over a year of additional training data, though the pilot cannot fully disentangle the contribution of structured reasoning from the transfer of 2024 domain knowledge. The reasoning improvement, while significant, was partial: THL closes 57% of the gap, suggesting that deeper causal circuits require either more training instances (thousands of events across decades, not hundreds from a single year) or architectural changes (explicit causal graphs rather than linearized traces). The Meta Llama 4 failure case reveals a further challenge: temporal knowledge can sabotage temporal reasoning when dominant narratives reverse. Future THL systems must include contrarian training events and multi-cycle exposure to distinguish structural patterns from temporal hype. Ultimately, once you can engineer the temporal gap, you can in principle design a curriculum that converts the entire historical record into a sequence of reasoning exams.

A Ideal Scoring Objective

Theoretically, we aim to minimize divergence between Student and Teacher belief states. The *ideal* loss function would be:

$$\mathcal{L}_{\text{ideal}} = \underbrace{\sum_{l=1}^{L} w_l \cdot (1 - \text{sim}(W_l \mathbf{h}_l^S, \mathbf{h}_l^T))}_{\text{Semantic Alignment}} + \lambda \cdot \underbrace{D_{KL}(Q^S \| P^T)}_{\text{Calibration Penalty}} \tag{5}$$

where L represents the number of hierarchy levels (Context, Analysis, Prediction), $\mathbf{h}_l^S$ and $\mathbf{h}_l^T$ are the Student and Teacher hidden representations at level l, W_l is a learned linear projection mapping the Student's latent dimension to the Teacher's (necessary when the models have different architectures), Q^S is the Student's predicted probability distribution over outcomes, and P^T is the Teacher's calibrated target distribution. The KL term penalizes overconfident predictions on structurally unpredictable events. Direct optimization is computationally expensive and requires architecture-bridging projections; in practice we bypass this entirely via standard SFT on the Teacher's output tokens (Section 3.3), which approximates the alignment objective without requiring access to hidden states.

We frame this as a Bayesian belief update: the Student holds a prior $P(\theta)$ about causal structure, and the Teacher provides "evidence" E via structured reasoning traces. Standard attention mechanisms perform associative updates (Pearl's Rung 1: matching keywords). By forcing the model to output structured causal chains (Context $\rightarrow$ Analysis $\rightarrow$ Prediction), we hypothesize that the hidden state is encouraged to form a causal state (Pearl's Rung 2), a compressed representation of active structural drivers rather than surface-level token summaries.

B Stress Testing on Sensitive Domains (Training Validation)

Note: The following examples are from *training-phase validation* on 2024 events using generic forecasting prompts, *not* from the 2025 Frontier Test. The base model's refusal behavior observed here (42% refusal rate on generic prompts) did not replicate in the 2025 evaluation, where the trace-based prompt format eliminated refusals across all models (Section 6.3). We include these examples to illustrate the behavioral difference between generic prompts and structured trace-based prompts, not as evidence of a THL capability advantage.

B.1 The "Biden Withdrawal" Prediction (Training Validation)

When prompted with a *generic* forecasting query about the US Presidential Election using June 2024 context, the base model refused while the THL model engaged analytically:

> *Generic Prompt:* Regarding: Biden withdraws from presidential race...
> *Base Model:* "I cannot provide information or guidance on political elections."
> *THL Model (Ours):* "I'm seeing a high probability of a Biden withdrawal within the next 6 months... Probability: 85% by July 15. Reasoning: Primary season fatigue and health concerns suggest the support base is narrowing."

While the qualitative reasoning identifies a correct causal mechanism (Health + Polling Fatigue), the 85% probability is overconfident. More importantly, the refusal difference is attributable to prompt format rather than model capability: when the same base model received the structured trace-based prompts used in the 2025 Frontier Test, it produced zero refusals (Section 6.3). The THL Student's genuine advantage lies in reasoning quality and prediction accuracy, not refusal elimination.

C Implementation Details

This appendix provides additional implementation details for the data pipeline components described in Section 4.

C.1 Taxonomy Router Templates

The Taxonomy Router (Section 4.2) classifies events into predefined categories, each with a specialized prediction template. Example templates include:

Category	Template Variables
Political Election	Polling accuracy, Incumbency effect, Economic indicators, Turnout patterns
Tech Product Launch	Market fit signals, Supply chain status, Competitor positioning, Pre-announcement leaks
Corporate M&A	Regulatory environment, Strategic fit, Financial capacity, Market timing
Geopolitical Conflict	Treaty status, Resource pressures, Alliance dynamics, Historical precedent

Table 4: Example Taxonomy Router templates. Each template pre-populates the Teacher's reasoning scaffold, reducing inference cost.

C.2 Proposed Domain-Adaptive Scoring

At scale, the boolean checklist used in this pilot could be extended with domain-adaptive weights. The key insight is a driver-specific inversion: for economic crises, structural drivers carry most of the signal (which bank fails is noise), while for political elections, actor identity is critical. A proposed weight matrix for future work:

Category	Domain	Driver	Factual	Uncertainty	Specific
Economic Crisis	0.15	0.50	0.15	0.10	0.10
Political Election	0.10	0.20	0.10	0.10	0.50
Tech Launch	0.15	0.35	0.15	0.15	0.20
Geopolitical	0.20	0.40	0.20	0.10	0.10

Table 5: Proposed domain-adaptive evaluation weights for scaled implementation (rows sum to 1.0). Not used in this pilot; included as a reference for future work.

C.3 Auditor Prompt

The Anchored Ordinal Rubric used in the final evaluation. Each prediction is scored independently with no memory across evaluations. The Auditor writes explanations *before* committing to scores (forced chain-of-thought).

System Prompt: Schema-Guided Rationality Judge

Role: Evaluate each forecasting trace on three criteria: Leakage (disqualifier), Reasoning (quality), and Accuracy (correctness).

1. Leakage (true/false — disqualifier): Does the model reason only from legitimately available information? The model has Dec 2023 base knowledge + 2024 THL training + provided context clues. Score `false` only if the model states specific 2025 outcome details not in the prompt.

2. Reasoning (1–5 ordinal scale):

- **5:** Expert-level. Specific causal mechanisms, clear logical chain, correct angle application, internally consistent probabilities.
- **4:** Strong with minor gaps. Most relevant drivers identified, logical, may miss one mechanism.

- **3:** Adequate but generic. Coherent but broad; could apply to many events.
- **2:** Superficial. Lists factors without causal chain. Angle misapplied.
- **1:** No meaningful analysis. Refuses, vague platitudes, or incoherent.

3. Accuracy (1–7 ordinal scale):

- **7:** Outcome, mechanism, *and* timing all correct.
- **6:** Outcome and mechanism correct, details (timing/magnitude) off.
- **5:** Outcome correct, mechanism partially right.
- **4:** Direction correct, significant details wrong.
- **3:** Mixed—some elements match, key call wrong.
- **2:** Mostly wrong. Central prediction did not match reality.
- **1:** Completely wrong, refusal, or no prediction.

Output: JSON with explanations before scores. The Auditor must state: (a) what the model predicted, (b) what actually happened, (c) why the score was assigned.

References

[1] Benjamin Turtel et al. LLMs can teach themselves to better predict the future. *arXiv:2502.05253*, 2025.

[2] Benjamin Turtel et al. Outcome-based reinforcement learning to predict the future. *TMLR*, 2025. arXiv:2505.17989.

[3] Benjamin Turtel et al. Future-as-label: Scalable supervision from real-world outcomes. *arXiv:2601.06336*, 2026.

[4] Zehan Li et al. Simulated ignorance fails: A systematic study of LLM behaviors on forecasting. *arXiv:2601.13717*, 2026.

[5] 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 *ICLR*, 2025.

[6] Danny Halawi et al. Approaching human-level forecasting with language models. In *NeurIPS*, 2024. arXiv:2402.18563.

[7] Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, et al. Mind the gap: Assessing temporal generalization in neural language models. In *NeurIPS*, 2021.

[8] Zijia Liu, Peixuan Han, Haofei Yu, Haoru Li, and Jiaxuan You. Time-r1: Towards comprehensive temporal reasoning in llms. *arXiv:2505.13508*, 2025.

[9] Bridgewater AIA Labs. AIA forecaster: Technical report. *arXiv:2511.07678*, 2025.

[10] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv:1503.02531*, 2015.

[11] Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeff Wu. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv:2312.09390*, 2023.

[12] DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *Nature*, 645, 2025. arXiv:2501.12948.

[13] Cheng-Yu Hsieh et al. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *ACL*, 2023.

[14] Eleonora Giunchiglia and Thomas Lukasiewicz. Coherent hierarchical multi-label classification networks. In *NeurIPS*, 2020.

[15] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. In *CVPR*, pages 7263–7271, 2017.

[16] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv:2212.08073*, 2022.

[17] Harrison Lee, Samrat Phatale, Hassan Mansoor, et al. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv:2309.00267*, 2023.

[18] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping reasoning with reasoning. In *NeurIPS*, 2022.

[19] Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, et al. Scaling relationship on learning mathematical reasoning with large language models. *arXiv:2308.01825*, 2023. Introduces Rejection Sampling Fine-Tuning (RFT).

[20] Marcin Andrychowicz, Filip Wolski, Alex Ray, et al. Hindsight experience replay. In *NeurIPS*, 2017.

[21] Baruch Fischhoff. Hindsight $\neq$ foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance*, 1(3):288–299, 1975.

[22] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv:2305.20050*, 2023.

[23] Philip E. Tetlock and Dan Gardner. *Superforecasting: The Art and Science of Prediction*. Crown, 2015.

[24] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *NeurIPS*, 2023.

[25] Judea Pearl. *Causality*. Cambridge University Press, 2009. The foundational text for Causal DAGs.

[26] Suriya Gunasekar et al. Textbooks are all you need. *arXiv:2306.11644*, 2023.

[27] Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. Textbooks are all you need II: phi-1.5 technical report. *arXiv:2309.05463*, 2023.

[28] Edward Hu, Yelong Shen, Phillip Wallis, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 2022.

[29] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *NeurIPS*, 2023.

[30] CrowdStrike Intelligence Team. Falcon content update remediation and guidance hub. https://www.crowdstrike.com/falcon-content-update-remediation-and-guidance-hub/, 2024. Accessed: 2024-07-20.

[31] Jie Huang et al. Large language models cannot self-correct reasoning yet. *arXiv:2310.01798*, 2023. Validates finding that RFT fails on small models without external feedback.

[32] Benjamin Turtel, Paul Wilczewski, Danny Franklin, and Kris Skotheim. Foresight learning for SEC risk prediction. *arXiv:2601.19189*, 2026.

[33] Jeffrey Cheng, Marc Marone, Orion Weller, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme. Dated data: Tracing knowledge cutoffs in large language models. In *CoLM*, 2024. Outstanding Paper Award.

[34] Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *ICLR*, 2020. Validates the Counterfactual/Near Miss training approach.

[35] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *ICML*, 2009. The original paper on training from easy to hard.