Future-as-Label: Scalable Supervision from Real-World Outcomes

Benjamin Turtel [1] **Paul Wilczewski** [1] **Danny Franklin** [1] **Kris Skothiem** [1]

Abstract

Time creates free supervision: forecasts about real-world events resolve to verifiable outcomes. The passage of time provides labels that require no annotation. To exploit this structure, we extend reinforcement learning with verifiable rewards to real-world prediction over time. We train language models to make probabilistic forecasts from causally masked information, using proper scoring rules as the reward function once events resolve. Learning is driven entirely by realized outcomes, enabling scalable outcome-based supervision in open-world prediction. On real-world forecasting benchmarks, Qwen3-32B trained using Foresight Learning improves Brier score by 27% and halves calibration error relative to its pretrained baseline, and outperforms Qwen3-235B on both constructed future-event prediction tasks and the Metaculus benchmark despite a 7× parameter disadvantage.

Training data: Hugging Face dataset

Model weights: Hugging Face model repo

Data generation: `https://lightningrod.ai/`

1. Introduction

Reinforcement learning with verifiable rewards has recently emerged as an effective approach for improving language models in domains such as mathematics, code generation, and formal reasoning, where correctness can be checked automatically. By replacing human annotation with deterministic reward functions, these methods scale efficiently and yield strong empirical gains. However, their applicability depends on the availability of immediate, closed-form verification, restricting them to tasks where correctness can be resolved at training time. As a result, despite their success, existing RLVR approaches remain confined to a narrow class of problems defined by readily available, task-specific reward signals.

In contrast, many real-world processes evolve over time and resolve to objective outcomes that are independent of the model. These outcomes, such as the conclusion of an election or the decision in a court case, are publicly observable and verifiable after the fact. This temporal structure induces a natural asymmetry between the information available at prediction time and the information revealed at resolution, creating a setting in which predictions can be evaluated retrospectively without relying on contemporaneous labels.

Our goal is to translate this temporal structure into a scalable learning framework for language models. Building on prior work on Foresight Learning (Turtel et al., 2025), we formalize learning from real-world temporal streams by grounding supervision in event resolution. The key constraint is causal: at prediction time t, the model is restricted to information available up to t, while evaluation is deferred until the corresponding outcome is realized. This formulation extends reinforcement learning with verifiable rewards beyond closed-world tasks with immediate feedback to settings where correctness is determined only after external, real-world resolution.

We adopt a reward-based objective that frames prediction as a stochastic decision evaluated retrospectively after outcome resolution. In contrast to supervised fine-tuning, which fits fixed targets, Foresight Learning optimizes over sampled reasoning trajectories using only outcome-based rewards, without intermediate annotations or task-specific labels. This perspective emphasizes calibration and decision quality rather than target matching. While we focus on binary outcomes for clarity, the formulation generalizes naturally to richer outcome spaces, such as continuous, multi-class, and free-text outcomes.

In this work, we formalize learning from temporally resolved real-world events as an extension of reinforcement learning with verifiable rewards, introduce an annotation-free algorithm for learning from delayed, outcome-based supervision, and show that this approach yields substantial improvements in calibration and predictive accuracy over strong pretrained baselines.

[1]Lightning Rod Labs. Correspondence to: Benjamin Turtel <ben@lightningrod.ai>.

Preprint. January 16, 2026.

2. Related Work

2.1. Reinforcement learning with verifiable rewards

Reinforcement learning with verifiable rewards (RLVR) has demonstrated strong results in domains with immediate, algorithmically checkable feedback, such as mathematics and programming (Wen et al., 2025) (Su et al., 2025). These settings typically involve short horizons and tightly scoped environments, which simplify credit assignment and reward attribution. Foresight Learning extends this paradigm to settings where outcomes resolve only after substantial temporal delay and outside the model's control.

2.2. LLM-based forecasting and static supervision

Recent work applies large language models to forecasting real-world events using prompting, retrieval, ensembling, and supervised fine-tuning over historical questions (Halawi et al., 2024). In particular, Halawi et al. generate multiple candidate reasoning–prediction pairs and then use realized outcomes (via Brier score) to select high-performing outputs for supervised fine-tuning. While outcome information is therefore used for offline data curation, this learning remains non-interactive. Foresight Learning differs by incorporating outcome resolution directly into the training loop as reinforcement signals.

2.3. Model-based judges and endogenous rewards

Another line of work uses language models as evaluators or judges to provide scalable feedback in settings where objective reward functions are unavailable (Liu et al., 2025b). Such approaches enable efficient supervision, but the resulting rewards are model-mediated and derived from re-evaluating the same information available to the predictor, which can propagate the biases and limitations of the judge model.

Foresight Learning differs in the source of supervision rather than the evaluation mechanism itself. While language models may assist in outcome resolution (e.g., assessing free-text evidence), the resolver has access to information that is causally unavailable at prediction time. Rewards are therefore grounded in externally resolved outcomes rather than alternative interpretations of the same input, ensuring that supervision reflects genuinely new evidence revealed over time.

3. Method

We consider settings where supervision is provided by the eventual resolution of events rather than contemporaneous labels. At prediction time t, the model observes only information available up to that cutoff and predicts whether an event will occur by a later time s >t. Although training is performed on events whose outcomes are already known, inputs are causally filtered to exclude post-t information, and rewards are computed solely from outcome resolution at time s, preserving the temporal asymmetry of prediction by construction.

3.1. Learning formulation

Each episode corresponds to a single future-event prediction.

Predictor and resolver roles.
Foresight Learning decomposes learning from temporal streams into two roles with asymmetric information access:

- The **predictor** is the language model being trained. At time t, it observes a temporally masked information state and produces a probabilistic prediction about a future event.

- The **resolver** is an external, fixed process that determines the realized outcome once the event resolves at time $s > t$. The resolver is implemented using a pretrained, frozen language model that is not trained, updated or influenced by the learning process. The resolver has access to post-t information unavailable to the predictor and is used solely to resolve outcomes, not score or rank predictions.

The predictor and resolver are strictly separated: the predictor never observes resolution information, and the resolver does not observe model outputs or training dynamics. Learning is driven by the **information gap** between the predictor's masked view at time t and the resolver's unmasked view at time s.

State.
The state consists of all information available up to time t, including relevant dated text and a natural-language specification of an event guaranteed to resolve by time $s > t$. The predictor operates under a masked information state, with all post-t information causally excluded by construction.

Action.
Conditioned on the state, the policy samples an internal reasoning trajectory terminates in a probabilistic prediction $p \in (0, 1)$, represented as a scalar value rather than a generated token. Only this numeric probability is exposed to the environment.. Formally, the action is the emitted probability; the trajectory is an internal stochastic computation optimized via policy gradients.

Reward.
Once the event resolves, a terminal reward is assigned using the log score:

$$\text{Reward} = y \cdot \log(p) + (1 - y) \cdot \log(1 - p)$$

where $y \in \{0, 1\}$ is the realized outcome. This strictly proper scoring rule incentivizes calibrated probabilistic predictions and provides a continuous learning signal under uncertainty.

Outcome determination is performed by a separate resolver that observes the unmasked future. The resolver has access to post-t sources unavailable to the predictor and is used solely to verify whether the event occurred. Each episode terminates after outcome resolution; there are no intermediate rewards.

Although the terminal reward takes the form of a proper scoring rule, this learning setup is not simply supervised likelihood training. In expectation, optimizing this reward corresponds to maximizing the log-likelihood of realized outcomes conditioned on the information available at prediction time. However, the learning problem is structured differently: the predictor acts under a causally masked information state without access to outcomes, and training optimizes a stochastic policy over reasoning trajectories whose quality is evaluated only after outcome resolution. Credit assignment is performed via policy gradients on sampled trajectories rather than by directly differentiating a likelihood objective, preserving the decision-theoretic structure of acting under asymmetric information.

This formulation treats prediction as a stochastic decision evaluated retrospectively after outcome resolution, distinguishing it from supervised likelihood training even though the reward takes the form of a proper scoring rule.

3.2. Objective and optimization

The objective is to maximize expected terminal reward under outcome-based supervision. In this regime, single-sample policy gradients exhibit high variance due to sparse terminal feedback and intrinsic uncertainty in event outcomes. To address this, we optimize using **Group Relative Policy Optimization (GRPO)**, as formulated by (Liu et al., 2025a).

For each state, the policy samples a group of K trajectories, each producing a probabilistic prediction. After outcome resolution, a reward is computed for each trajectory. We define a group-relative advantage by subtracting the mean reward within the group:

$$\text{Advantage}(\tau_i) = \text{Reward}(\tau_i) - \frac{1}{K} \sum_j \text{Reward}(\tau_j)$$

Policy updates maximize the expected advantage-weighted log probability of each trajectory under the current policy. By comparing trajectories generated under identical pre-t information, GRPO reduces variance from outcome noise and stabilizes learning when supervision is provided only through terminal outcomes. Gradients are applied to all tokens in each trajectory, enabling credit assignment across extended reasoning processes even though feedback is available only at the final step.

3.3. Training protocol

We explicitly enforce a causal information constraint by applying a temporal information mask to the input stream. For each prediction time t, the predictor is restricted to observing only information timestamped at or before t, even though training is performed offline. All training events resolve strictly after the pretrained model's knowledge cutoff, ensuring that realized outcomes cannot be encoded in the model's parametric memory. All post-t information - including sources required to determine the outcome - is withheld during prediction and policy optimization. Outcome verification is performed by a separate resolver with access to the unmasked stream. This preserves a strict causal separation between observation, action, and verification throughout training.

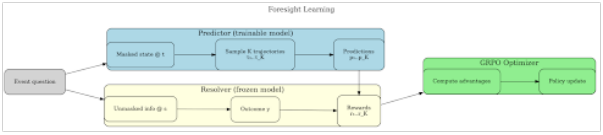

Figure 1. Overview of Foresight Learning

4. Experimental Setup

4.1. Dataset construction

We construct a future-event prediction dataset designed to preserve a strict temporal separation between prediction and verification. The pre-cutoff information state consists of an English-language news corpus aggregated from publicly accessible outlets (e.g., international newspapers, wire services, and financial news sites). Articles are timestamped using publisher-provided publication times, normalized to UTC.

For each example, we freeze the news corpus at a cutoff time t and generate a binary question about an event expected to resolve strictly after that cutoff, using only information available prior to t. The cutoff is defined with respect to publisher-reported publication timestamps; articles with missing or ambiguous timestamps are excluded to prevent temporal leakage. Generated events span multiple domains,