including politics, economics, and corporate actions.

To prevent information leakage, model inputs are constructed exclusively from sources published at or before time t. Outcome verification uses independent post-t sources that are not included in the model's input context. Event outcomes are resolved automatically by a separate, frozen large language model (Gemini-2.5-Flash) with access to a broader pool of post-cutoff news and archival sources, and used solely to determine whether an event occurred. The resolver does not observe model outputs or training dynamics; as a result, resolution errors introduce noise but do not induce endogenous reward signals. Examples that cannot be resolved with high confidence are discarded. Each event is assigned a resolution time s, defined as the earliest dated source supporting the resolved outcome.

All questions and outcomes are generated prior to training, enabling fully offline optimization while preserving the temporal and causal structure of real-world prediction.

4.1.1. Dataset Statistics and Splits

The full dataset contains 5,620 binary prediction examples. Of these, 5,120 examples are used for training, and 500 examples are held out as a temporally disjoint test set constructed using the same event-generation procedure. In addition, we evaluate on a second, independent test set consisting of 293 human-written forecasting questions from Metaculus, which are never used during training or data construction.

We intentionally do not construct a validation set: models are trained on all available pre-test data using a fixed training procedure, without early stopping or model selection based on held-out examples. Training data consists of predictions made as of July 1, 2024 through January 30, 2025. Both test sets consist exclusively of predictions made on or after February 1, 2025, ensuring strict temporal separation between training and evaluation.

4.1.2. Task characteristics

Prediction horizons range from days to several weeks, allowing learning across varying outcome horizons. Although outcomes are discrete, supervision and evaluation are based on continuous probabilistic scores, enabling analysis of accuracy and calibration under increasing temporal uncertainty.

4.2. Models and training

We fine-tune a Qwen3-32B language model with explicit reasoning enabled. Conditioned on an information state, the model generates a reasoning trajectory that terminates in a probabilistic prediction expressed explicitly at the end of the output. Parsed probabilities are constrained to the interval [0.001, 0.999] for numerical stability.

Training is performed using GRPO. For each event, the model samples four independent trajectories, each producing a probabilistic prediction. After outcome resolution, a log-score reward is computed for each trajectory, and relative advantages are obtained by subtracting the per-group mean reward. Policy updates increase the relative likelihood of higher-reward trajectories. Training uses batches of 32 events, with prediction horizons mixed within each batch.

4.2.1. Baselines

We compare Foresight Learning to baselines that operate under identical temporal constraints and produce probabilistic predictions in the same output format, isolating the effect of learning.

Prompted forecasting: the base Qwen3-32B and Qwen-235B models are prompted to produce probabilistic predictions without task-specific fine-tuning. This baseline measures forecasting performance without learning from outcome resolution.

Ensembling: multiple independent predictions are generated per event and averaged to assess gains from sampling and aggregation without parameter updates. This control tests whether improvements can be explained by variance reduction alone.

4.2.2. Evaluation metrics

We evaluate models based on the quality of probabilistic predictions. We report the **log score** used for training, the **Brier score**, which measures squared error between predicted probabilities and outcomes, and **calibration**, assessed via expected calibration error (ECE) over 10 discretized probability bins measuring empirical outcome frequencies as a function of predicted confidence.

5. Results

We evaluate Foresight Learning on two held-out test sets: (i) a synthetic future-event benchmark of 500 questions constructed under strict temporal controls, and (ii) an external benchmark consisting of 293 binary forecasting questions from Metaculus. Performance is evaluated using proper scoring rules and calibration metrics.

Table 1 compares four inference regimes: (i) Qwen3-32B prompted for a single forecast, (ii) Qwen3-32B prompted for seven independent forecasts with the median taken as the final prediction, (iii) Qwen3-235B prompted for a single forecast, and (iv) the Foresight-trained model prompted once. Repeated prompting and median aggregation provide modest improvements over single-sample prompting but do not match the gains from training on resolved outcomes. Notably, the Foresight-trained 32B model outperforms both the

Table 1. Forecasting performance on synthetic and real-world benchmarks

Model	Log ↑	Brier ↓	ECE ↓
Metaculus			
Qwen3-32B	-0.7210	0.2472	0.2175
Qwen3-32B Ensemble	-0.7000	0.2390	0.2289
Qwen3-32B-RL (160)	**-0.5738**	**0.1793**	**0.1042**
Qwen3-235B	-0.6828	0.2111	0.1905
Synthetic future-events			
Qwen3-32B	-0.7166	0.2432	0.1732
Qwen3-32B Ensemble	-0.7045	0.2481	0.1864
Qwen3-32B-RL (160)	**-0.5978**	**0.1979**	**0.0598**
Qwen3-235B	-0.7138	0.2260	0.1695

ensemble-style baseline and the substantially larger 235B model across all metrics, indicating that the improvements stem from the training objective rather than increased sampling or model scale.

Performance gains persist on the Metaculus benchmark, which consists of independently authored questions outside the synthetic benchmark distribution. One possible contributing factor is that Metaculus questions often concern higher-salience events with broader public coverage, providing richer information at prediction time. While this hypothesis requires further study, the results indicate that learning from externally resolved outcomes generalizes beyond the specific data construction process used for training.

Taken together, these results support the central premise of Foresight Learning: incorporating outcome resolution directly into the training objective yields more accurate and better-calibrated probabilistic forecasts than prompting or sampling-based baselines alone, even when compared to substantially larger pretrained models.

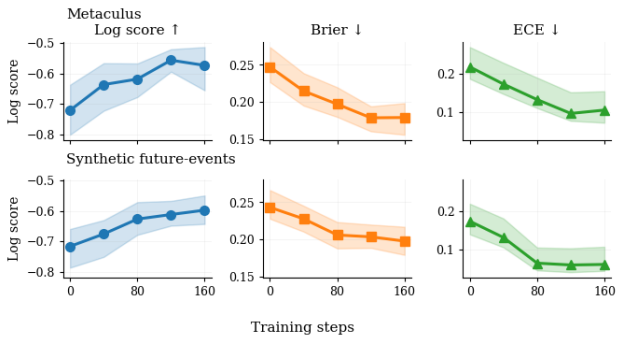

Figure 2. Model calibration and accuracy metrics versus training steps on Metaculus (top) and synthetic future-events (bottom). Shaded regions show 95% bootstrap confidence intervals. Metrics are log score (↑), Brier score (↓), and expected calibration error (ECE; ↓). Performance improves monotonically with training.

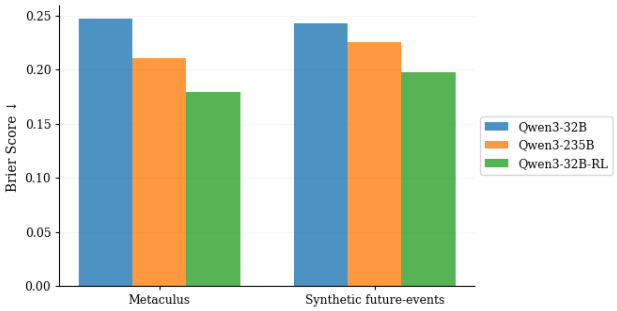

Figure 3. Brier scores (↓) for different models on Metaculus and synthetic future-events benchmarks. Foresight Learning consistently outperforms both the base and larger pretrained baselines.

6. Discussion and Conclusion

This work studies a supervision regime in which feedback is provided by the eventual resolution of real-world events rather than contemporaneous labels or proxy objectives. By optimizing probabilistic predictions retrospectively using proper scoring rules, Foresight Learning aligns training with the temporal and causal structure of forecasting under uncertainty.

Empirically, learning from outcome resolution improves probabilistic forecasting performance relative to a strong pretrained baseline, with consistent gains in accuracy and calibration on both synthetic future-event datasets and the independently authored Metaculus benchmark. Notably, Foresight Learning materially outperforms a substantially larger same-generation model on real-world forecasting tasks.

A key benefit of outcome-based supervision is improved calibration. Because rewards are assigned only after outcomes resolve, overconfident incorrect predictions incur large penalties, while appropriately uncertain predictions are penalized less severely. This learning signal encourages inference strategies that balance evidence aggregation with uncertainty estimation, whereas sampling-based heuristics such as ensembling reduce variance without modifying the underlying prediction policy.

Relative to prior reinforcement learning with verifiable rewards, Foresight Learning operates in open-world domains with sparse and delayed feedback. Trajectory-level, group-relative optimization enables stable credit assignment under long and variable horizons by comparing alternative predictions generated under identical informational constraints and evaluating them retrospectively after outcomes resolve.

This work has several limitations. Training is performed offline on resolved events; while deployment-time feedback loops are under active exploration, they are not evaluated

in this study. Event specification and outcome resolution rely on automated pipelines that may introduce biases or coverage gaps, and the current experiments focus on binary outcomes. Extending the framework to richer outcome spaces and fully online settings remains an important direction for future work.

Overall, Foresight Learning demonstrates that effective supervision can arise directly from chronologically evolving real-world data. By incorporating outcome resolution into the training objective, the framework points toward a broader role for outcome-based supervision in extending verifiable reward-driven learning beyond closed-world tasks and toward open-ended, real-world decision-making.

7. Data and Model Availability

The trained model, datasets, and data generation platform are publicly available to support reproducibility and future research.

Training data: Hugging Face dataset

Model weights: Hugging Face model repo

Data generation: https://lightningrod.ai/

References

Halawi, D., Zhang, F., Yueh-Han, C., and Steinhardt, J. Approaching human-level forecasting with language models, 2024. URL https://arxiv.org/abs/2402.18563.

Liu, S., Liu, H., Liu, J., Xiao, L., Gao, S., Lyu, C., Gu, Y., Zhang, W., Wong, D. F., Zhang, S., and Chen, K. Compassverifier: A unified and robust verifier for llms evaluation and outcome reward, 2025a. URL https://arxiv.org/abs/2508.03686.

Liu, Z., Chen, C., Li, W., Qi, P., Pang, T., Du, C., Lee, W. S., and Lin, M. Understanding r1-zero-like training: A critical perspective, 2025b. URL https://arxiv.org/abs/2503.20783.

Su, Y., Yu, D., Song, L., Li, J., Mi, H., Tu, Z., Zhang, M., and Yu, D. Expanding rl with verifiable rewards across diverse domains, 2025. URL https://arxiv.org/pdf/2503.23829.

Turtel, B., Franklin, D., Skotheim, K., Hewitt, L., and Schoenegger, P. Outcome-based reinforcement learning to predict the future, 2025. URL https://arxiv.org/html/2505.17989v4.

Wen, X., Liu, Z., Zheng, S., Ye, S., Wu, Z., Wang, Y., Xu, Z., Liang, X., Li, J., Miao, Z., Bian, J., and Yang, M. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms, 2025. URL https://arxiv.org/abs/2506.14245.