The Oracles

commit fb221514

ForecastingPath probabilistic forecasting endpoint

Team CanadaHacks - Prophet Hacks 2026

Single-binary Brier

0.0378

lower is better

95% paired bootstrap CI

[0.0143, 0.0374]

Opus 4.7 vs Sonnet 4.6, n=26

Primary discipline finding

85% floor fix

safety-net bug > model swap

What runs live

- FastAPI endpoint on Railway: POST /predict returns per-outcome probabilities.
- Pipeline: event JSON, Brave evidence, source ranking, Opus 4.7 forecast, Kalshi longshot floor, structured JSON.
- Every prediction writes a private trace: query, evidence URLs, raw model output, parser path, latency, and warnings.
- Public root stays sparse; dashboard, reports, galleries, and traces are auth-gated during active scoring.

Current caveats

- Resolved-set retrieval leaks future information on some rows; absolute backtest Brier is best-case-with-hindsight.
- PA CLI single-binary and proper multi-class Brier rank variants differently; claims must name the metric.
- First live Prophet Arena call is still the trigger for payload-shape, latency, and scoring validation.

Measured variants

Same retrieval and prompt; only model call changes.

Variant	Binary	Multi
Claude Opus 4.7 (prod)	0.0378	0.2558
Claude Opus 4.6	0.0391	0.2500
OpenAI GPT-5.2	0.0438	0.2874
Claude Sonnet 4.6	0.0639	(0.6912)
OpenAI GPT-5.5	0.0920	0.3429
Gemini 3.1 Pro	0.0983	0.4773

Limitations

- Live performance is unverified. Prophet Arena scoring starts after submission.
- The 0.0378 backtest used resolved events; 38.5% of evidence URLs contain post-resolution markers. Live numbers will likely be higher.
- Sample is small (n=26) and 62% sports. A balanced eval would widen the picture.

Demo path

Public root -> dashboard -> observatory -> pipeline trace -> abstain slider -> heatmap -> review brief

Generated from submission/REPORT.md by scripts/build_submission_onepager.py