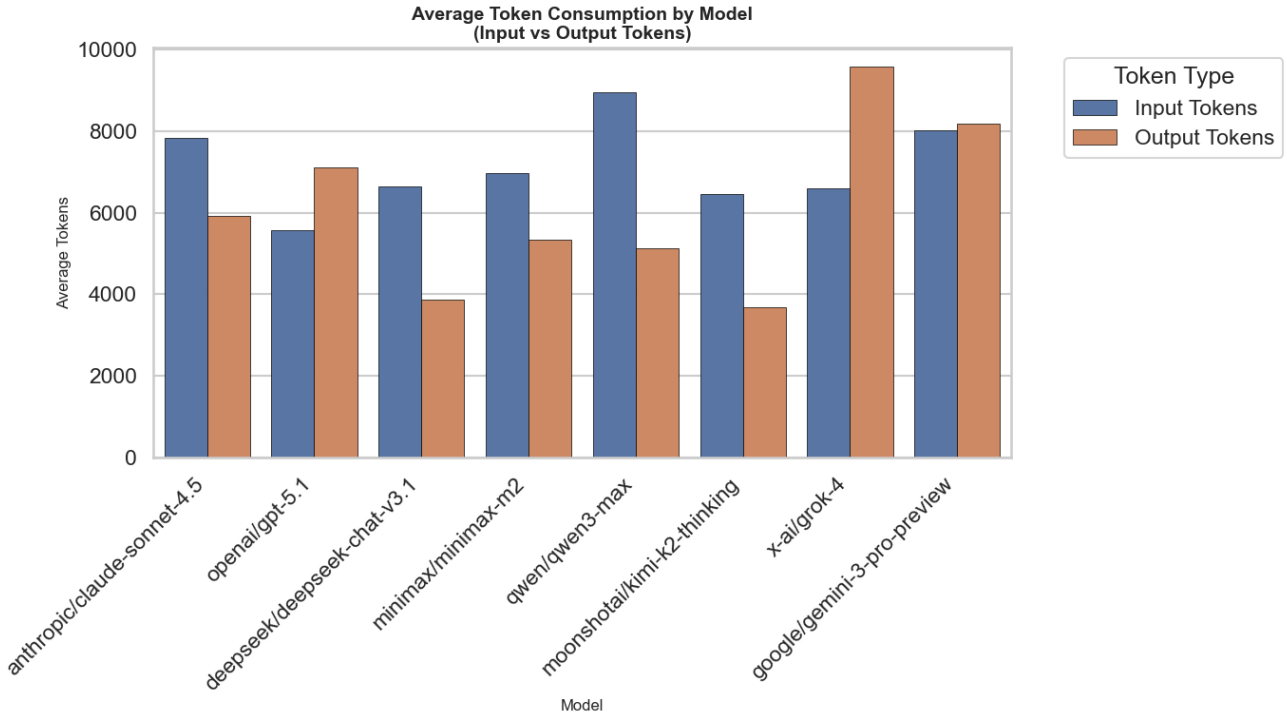

Figure 4. Average input and output token consumption by model.

5. Benchmarking and Results

TruthTensor introduces a benchmarking framework that evaluates large language models in dynamic, real-world forecasting environments rather than static datasets. The benchmark is designed to measure probabilistic reasoning, temporal consistency, and human imitation fidelity under uncertainty.

5.1. Benchmark Design Principles. The benchmarking framework is guided by the following principles:

- Market grounding: All benchmarks are anchored to real prediction markets with externally resolved outcomes.
- Temporal evaluation: Models are evaluated longitudinally rather than through single-shot predictions.
- Prompt immutability: Instruction locking prevents prompt optimization from influencing benchmark outcomes.
- Model-agnosticism: The benchmark supports frontier, mid-tier, and distilled models under identical conditions.
- Drift awareness: Benchmarks explicitly measure forecast drift and update behavior over time.

These principles ensure that benchmarking reflects realistic decision-making conditions and mitigates common sources of evaluation leakage.

5.2. Benchmark Tasks. Benchmark tasks consist of forecasting questions drawn from live prediction markets. Tasks vary across domains, risk profiles, and temporal horizons, enabling broad coverage of reasoning scenarios. These include binary outcome markets, multinomial outcome markets, short-horizon events with rapid information updates, and long-horizon events with sparse information flow.

Each task on TruthTensor is evaluated from market inception to resolution, capturing the full lifecycle of belief formation and revision. The benchmark is conducted on live prediction markets over a 30-day window (Dec 12, 2025 – Jan 10, 2026). During this period the platform processed 876,567 forecasting decisions, with 531,770 users, 983,600 fine-tuned agents, and over $1.14B in active market value. In total, more than 1.18M probability updates were recorded across active markets spanning geopolitics, economics, science, and technology. This evaluation differs from static test sets in that models operate concurrently on the same markets at the same timestamps, enabling paired comparisons under identical information conditions.

5.3. Compared Models. The benchmark evaluates multiple classes of models under identical configurations, frontier proprietary LLMs, open-weight large language models, distilled or parameter-efficient models, and baseline forecasting strategies. All models operate under the same instruction templates, token budgets, and information access constraints to ensure comparability. In total, eight large-scale models meet the inclusion criteria of (i) at least 50,000 deployed agents and (ii) sustained live forecasting activity across the full evaluation window. These eight models account for essentially all benchmark activity during the evaluation period. Models with negligible deployment were excluded to avoid small-sample bias. See Table 2 for details.

5.4. Baseline Benchmarks. TruthTensor includes multiple baseline benchmarks to contextualize model performance. These include market-implied probability baseline, uniform probability baseline, historical frequency baseline, and simple heuristic baselines. These are evaluated using the same metrics and time schedules as LLM-based agents.

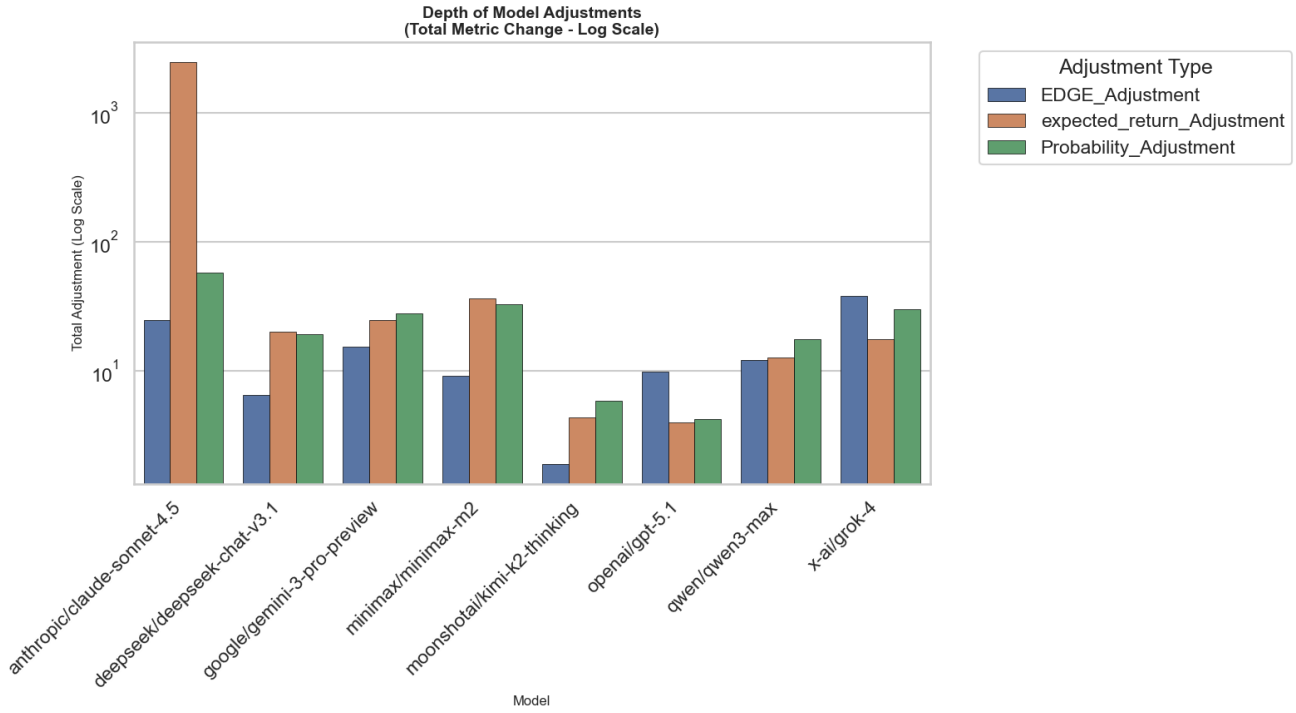

Figure 5. Depth of model adjustments across edge, expected return, and probability dimensions (log scale).

5.5. Benchmark Metrics. Benchmark performance is assessed using the evaluation metrics defined in Section 4. Metrics are reported both per-market and aggregated across markets and categories. These include probabilistic accuracy metrics, calibration metrics, drift and temporal coherence metrics, market alignment metrics, and risk-adjusted performance metrics. This multi-dimensional reporting prevents over-optimization for any single metric.

Because all models forecast the same events at the same times, differences in Brier score, calibration, and drift reflect genuine differences in probabilistic reasoning and temporal updating rather than dataset artifacts. Models with low Brier but high drift exhibit unstable narratives (frequent probability swings beyond what is justified by market movement), while models with low ECE but high mean absolute deviation between their predicted probability and the market's probability, demonstrate systematic disagreement with collective belief.

5.6. Benchmark Aggregation. Benchmark results are aggregated across event domains, risk categories, temporal horizons, and market liquidity levels. Aggregation enables identification of systematic strengths and weaknesses across models rather than performance on isolated tasks.

5.7. Reproducibility and Reporting. All benchmarking runs log model versions, prompt hashes, timestamps, and evaluation outputs. Completed markets are included without filtering. Benchmark definitions and metric calculations remain fixed throughout evaluation to ensure reproducibility. The benchmarking framework is designed to be extensible, allowing new markets, models, and baselines to be incorporated without altering existing results.

5.8. Behavioral and Temporal Diagnostics. Figure 2 illustrates both the adjusted cumulative performance of the model relative to historical baselines (left) and the realized cumulative trading profit and loss over time during the TruthTensor experiment (right). The left panel demonstrates that the proposed agent consistently outperforms market, uniform, and historical benchmark strategies, exhibiting a monotonic upward trajectory that reflects stable probabilistic calibration and sustained positive edge extraction across evaluation rounds. In contrast, the right panel captures realized market exposure under the fixed 30-position portfolio constraint, highlighting periods of volatility in which some strategies incur drawdowns due to miscalibration, drift amplification, or high-risk event exposure. The divergence between the adjusted and realized curves emphasizes the importance of separating intrinsic forecasting quality from market-execution noise. Notably, the dominant model maintains controlled losses during turbulent intervals and recovers steadily, whereas weaker baselines exhibit prolonged negative P&L, validating the robustness of the drift-aware, calibration-adjusted, and risk-constrained decision framework.

Figure 3 shows how frequently each model selects different decision strategies. Models such as GPT-5.1 and Grok-4 exhibit a heavy concentration in drift-adjusted strategies, indicating frequent recalibration to market movement. In contrast, models with a higher proportion of HOLD and UNKNOWN actions tend to behave more conservatively, trading off responsiveness for stability. This distribution explains much of the observed temporal drift: models that shift strategies aggressively also produce larger probability swings.

Figure 4 highlights substantial differences in reasoning footprint. Gemini-3-Pro-Preview and Grok-4 consume the largest

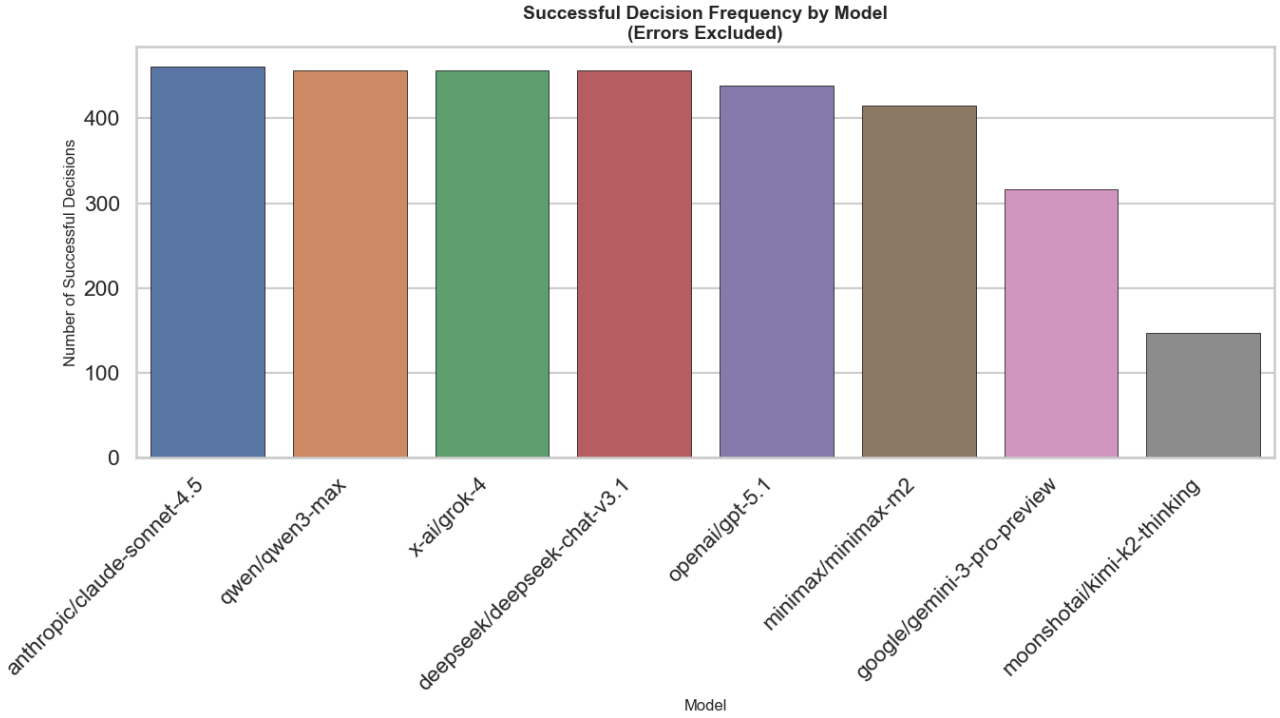

Figure 6. Frequency of successful decisions by model (errors excluded).

number of output tokens, reflecting long internal reasoning and explicit probability articulation. Models with smaller output budgets (e.g., Qwen3-Max, DeepSeek-Chat-v3.1) produce more compact updates, which empirically leads to lower variance in probability revisions. Larger token budgets enable richer causal modeling but also amplify sensitivity to transient information, increasing temporal drift.

Figure 5 shows how aggressively each model modifies its internal belief state. Claude-Sonnet-4.5 performs extremely large expected-return adjustments, indicating strong re-weighting of outcomes when new information arrives. Kimi-K2-Thinking exhibits minimal adjustment depth, producing more inert belief trajectories. These differences directly map to the drift metric: deeper adjustment profiles correspond to larger and more frequent probability shifts.

Figure 6 measures the frequency with which models successfully return valid system-prompt outputs within predefined decision-time and latency constraints. Claude-Sonnet-4.5, Grok-4, and DeepSeek-Chat-v3.1 achieve the highest success counts, indicating reliable local updates. Kimi-K2-Thinking, despite its long reasoning chains, exhibits the lowest success frequency, showing that verbosity does not necessarily translate into effective probabilistic correction. These diagnostics reveal that model quality in live forecasting is governed not only by final accuracy but by how beliefs evolve. High-capacity models generate stronger raw signals but also exhibit greater volatility and adjustment depth. More compact models trade expressive power for temporal coherence, yielding steadier probability trajectories. TruthTensor makes these behavioral differences measurable, allowing models to be compared not just by what they predict, but by how they reason over time.

6. Conclusion

This work presents TruthTensor as a new class of benchmark for evaluating large language models in forward-looking decision environments. By embedding LLM agents directly into live prediction markets, the framework observes not only what models predict, but how they revise their beliefs over time under real uncertainty. Across nearly one million probability updates and eight simultaneously deployed frontier-scale models, we show that performance cannot be summarized by a single score.

The results highlight a fundamental trade-off between accuracy, stability, and operational efficiency. High-capacity models provide deeper reasoning and often better outcome prediction, but at the cost of greater volatility and higher inference expense. Conversely, low-cost models offer stable, scalable forecasting but sacrifice expressive power and nuance. By quantifying these dimensions jointly, TruthTensor enables principled model selection for real-world forecasting tasks, where coherent belief trajectories and reliable uncertainty estimates are as important as final accuracy.

Beyond individual model comparison, TruthTensor demonstrates that continuous, live-forward evaluation is both feasible and informative at scale. The platform's ability to collect paired forecasts across models on identical events enables statistically robust benchmarking under conditions that closely mirror deployment. This opens the door to new forms of evaluation that track how models learn, adapt, and sometimes overreact in the face of unfolding information.

As large language models are increasingly used to support high-stakes decisions, benchmarks must move beyond static question answering toward measuring dynamic belief forma-