TRUTHTENSOR: EVALUATING LLMS THROUGH HUMAN IMITATION ON PREDICTION MARKET UNDER DRIFT AND HOLISTIC REASONING

Shirin Shahabi, Spencer Graham, and Haruna Isah
[shirin, spencer, and haruna]@inferencelabs.com
Inference Labs Inc.

Abstract

Evaluating language models and AI agents remains fundamentally challenging because static benchmarks fail to capture real-world uncertainty, distribution shift, and the gap between isolated task accuracy and human-aligned decision-making under evolving conditions. This paper introduces TruthTensor, a novel, reproducible evaluation paradigm that measures reasoning models not only as prediction engines but as human-imitation systems operating in socially-grounded, high-entropy environments. Building on forward-looking, contamination-free tasks, our framework anchors evaluation to live prediction markets and combines probabilistic scoring to provide a holistic view of model behavior. TruthTensor complements traditional correctness metrics with drift-centric diagnostics and explicit robustness checks for reproducibility. It specify human vs. automated evaluation roles, annotation protocols, and statistical testing procedures to ensure interpretability and replicability of results. In experiments across 500+ real markets (political, economic, cultural, technological), TruthTensor demonstrates that models with similar forecast accuracy can diverge markedly in calibration, drift, and risk-sensitivity, underscoring the need to evaluate models along multiple axes (accuracy, calibration, narrative stability, cost, and resource efficiency). TruthTensor therefore operationalizes modern evaluation best practices, clear hypothesis framing, careful metric selection, transparent compute/cost reporting, human-in-the-loop validation, and open, versioned evaluation contracts, to produce defensible assessments of LLMs in real-world decision contexts. We publicly released TruthTensor at https://truthtensor.com.

1. Introduction

Artificial intelligence (AI) evaluation has traditionally been anchored in static metrics and benchmark datasets designed to assess models' performance in predefined tasks, such as general problem-solving [1], code generation [2], or mathematical reasoning [3]. The fundamental issue with these traditional benchmarks is their reliance on closed-world evaluations, where AI models are tested on a fixed set of tasks or datasets that often contain historical information or established patterns. In these settings, models can perform exceptionally well by memorizing data or applying pattern-matching techniques, but these metrics fail to capture a model's ability to reason, adapt, or learn in novel contexts. As AI systems grow more complex and exhibit behaviors that seem to mimic human cognition, such as abstract reasoning, dynamic problem-solving, and real-time interaction, the static nature of these benchmarks becomes increasingly inadequate for assessing the true breadth and depth of AI capabilities.

In recent years, the AI community has acknowledged the limitations of these traditional evaluation frameworks. The reliance on outdated benchmarks for specialized tasks, while still necessary for certain applications, no longer suffices for evaluating the full range of cognitive abilities that next-generation models exhibit [4]. This gap is highlighted by initiatives like Chatbot arena[5], CORE-Bench [6], Futurex [7] and Humanity's Last Exam [8], which attempts to assess models' capabilities in more comprehensive and real-world contexts. However, even such ambitious projects are often limited by the underlying problem that they still measure how well models parse or reproduce knowledge derived from historical datasets, rather than how they might reason, extrapolate, or adapt to new, unseen scenarios.

1.1. TruthTensor: A Paradigm Shift from Prediction to Human Imitation. In response to the limitations of traditional AI evaluation frameworks, this work introduces TruthTensor, a novel approach that fundamentally reframes the evaluation question. Rather than asking "How well can this model predict future events?", a question that conflates memorization, pattern matching, and genuine reasoning, TruthTensor asks: "How well does this model imitate human reasoning, calibration, and narrative coherence when confronted with evolving, uncertain, socially-grounded scenarios?". This shift from prediction evaluation to human imitation assessment represents a fundamental departure from static evaluation methods that test information retrieval from historical datasets.

TruthTensor integrates LLMs with streaming real-time prediction market data, challenging their ability to reason through complex, high-entropy environments where ground truth has yet to be established. The framework measures not only whether models produce accurate forecasts, but how their reasoning processes, confidence calibration, and narrative stability compare to human market participants. Anchoring evaluation to prediction markets which aggregate financially backed, continuously updated crowd expectations, provides TruthTensor with a dynamic, socially grounded reference point for assessing LLMs human imitation capabilities.

1.2. Key Differentiators. TruthTensor differs from existing benchmarks in several critical ways:

- Human Imitation Focus: Evaluation centers on how well LLMs replicate human reasoning patterns, calibration, and narrative coherence, not merely prediction accuracy.
- Drift-Centric Design: The framework places primary emphasis on measuring narrative drift, temporal inconsistency, and reasoning confidence decay, dimensions largely ignored by existing benchmarks.
- Contamination-Free Construction: Evaluates only forward-looking events, thereby eliminating data contamination by construction, a fundamental weakness of static benchmarks.
- Baseline Independence: Baseline models provide reference points independent of rolling-window calibration, ensuring fair comparison across models with different training histories.
- Instruction Locking: Prompt specifications are versioned and locked, ensuring reproducibility and preventing prompt engineering from masking model limitations.
- Holistic Evaluation: Metrics span correctness, risk assessment, temporal coherence, calibration, and drift magnitude, providing a comprehensive view of model capabilities.
- Specified Evaluation Categories: Events are categorized by risk profile, domain, temporal horizon, and market liquidity, enabling targeted analysis of model strengths and weaknesses.

Together, these design choices motivate the background concepts and methodological foundations introduced in the following section.

2. Background

LLM evaluation extends far beyond assigning scores on fixed benchmarks. Evaluation serves multiple purposes including comparison, diagnosis, selection, and deployment assurance. Its design is inherently shaped by who is evaluating (model builders, researchers, or end users) and why it is carried out. The Evaluation Guidebook [9], highlights that no single benchmark or metric can fully characterize model quality as evaluations encode assumptions about tasks, data, prompts, and metrics, each introducing trade-offs and biases. The guidebook emphasizes the limitations of static leaderboards, the risks of overfitting to benchmarks, and the growing importance of custom, task-specific, and human-aligned evaluations. It further distinguishes between offline and online evaluation, automatic and human-in-the-loop methods, and capability versus behavior testing, while underscoring challenges such as data contamination, prompt sensitivity, and distributional shift. The guidebook motivates a more principled, transparent, and purpose-driven approach to evaluation, one that recognizes evaluation as an evolving system rather than a fixed score, providing the conceptual backdrop for alternative frameworks that prioritize robustness, interpretability, and real-world relevance.

2.1. Reasoning Beyond Training Data: Avoiding Contamination. Static evaluation on held-out test sets has been shown to overstate model performance. As training corpora grow ever larger, it becomes difficult to ensure that no part of a static benchmark has leaked into a model's training data. Recent analyses find that evaluating LLMs on fixed benchmarks is

vulnerable to data contamination and leaderboard overfitting, and that traditional test sets are inevitably affected by possible data contamination [4]. Forecasting-based evaluation avoids this pitfall by construction: models must predict future events that by definition did not exist during training [10].

TruthTensor extends this principle by exclusively evaluating forward-looking events whose outcomes are unknown at prediction time. This design choice ensures that no model can leverage memorized information about test set outcomes, forcing evaluation to focus on genuine reasoning capabilities rather than training data recall. The framework further mitigates contamination risk through instruction locking: prompt templates are versioned and immutable, preventing researchers from inadvertently introducing test set information through prompt engineering.

2.2. LLMs as Oracles: The Human Imitation Perspective. A growing body of research explores the use of LLMs as Oracles [11], systems that provide probabilistic assessments or judgments about uncertain events. This perspective reframes LLMs not as deterministic answer generators, but as probabilistic reasoning systems that, like human experts, must navigate uncertainty, update beliefs, and express confidence appropriately. TruthTensor adopts this oracle perspective, evaluating LLMs' ability to function as human-like probabilistic reasoners in socially-grounded contexts.

The oracle framework is particularly relevant for prediction markets, where human participants continuously update their probability estimates based on new information, market dynamics, and narrative shifts. Through the comparison of LLM outputs to market-implied probabilities which represent aggregated human expectations, TruthTensor measures how well models replicate human-like reasoning patterns, calibration, and narrative coherence. This comparison reveals not just prediction accuracy, but the quality of the underlying reasoning process: models that exhibit human-like drift patterns, appropriate confidence calibration, and narrative stability are better oracles than those that produce static, overconfident, or narratively inconsistent outputs.

2.3. LLM Evaluation and Benchmarks. New frameworks for evaluating large language models (LLMs) are moving away from one-size-fits-all benchmarks toward a combination of task-specific evaluations, human-centric assessments, and robustness tests [12]. Such frameworks consider the variety of ways that AI models are integrated into the real world, from their ability to understand complex instructions and engage in nuanced dialogue, to their capacity for handling ambiguous inputs and generating creative outputs.

The longstanding use of benchmarks like HumanEval [2][13] for coding tasks, GSM8K [3] for mathematical reasoning, and ARC [1] for general knowledge questions has been instrumental in providing an initial gauge for model performance. These tools, while useful for tracking progress over time, have increasingly been shown to offer a limited and outdated measure of the true cognitive capabilities of modern AI systems. Emerging discussions around AI evaluation have called for a shift toward open-world evaluation, where models are assessed on their ability to generalize, handle uncertainty, and demonstrate adaptive reasoning in ever-changing environments [14]. This approach aims to move beyond rote knowledge retrieval and instead measure how AI systems engage with novel problem-solving scenarios that require creativity, abstraction, and domain-independent thinking.

Recent research underscores the need for a more diverse and multi-dimensional approach to AI evaluation. For instance, the Chatbot Arena framework [5] addresses the limitations of conventional single-turn evaluation of conversational AI by introducing a multi-turn, interactive benchmarking environment. In this setup, multiple conversational agents, or agents interacting with human interlocutors, engage in structured dialogues across diverse scenarios. The resulting interactions are assessed along metrics such as coherence, consistency, helpfulness, and overall conversational quality.

Humanity's Last Exam [8] proposes a comprehensive evaluation framework that challenges large language models (LLMs) with a broad suite of tasks designed to assess general intelligence across multiple cognitive domains. Rather than focusing narrowly on individual capabilities, such as code generation or math reasoning, the benchmark includes diverse tasks spanning logic, commonsense reasoning, real-world knowledge, and abstract problem-solving. The design of Humanity's Last Exam emphasizes generalization and adaptability: models are evaluated on their ability to apply reasoning to new, previously unseen tasks, thereby reducing the likelihood that success stems merely from memorizing patterns in training data or overfitting to narrow benchmarks.

Futurex [7] introduces a dynamic forecasting benchmark designed to evaluate AI models' ability to reason and make probabilistic predictions about future events under uncertainty. Unlike traditional static datasets, all tasks in Futurex are forward-looking, with outcomes unknown at the time of prediction, effectively preventing data contamination and memorization. The benchmark spans geopolitical, economic, and social events, requiring models to produce probabilistic forecasts evaluated via proper scoring rules such as the Brier score and log-likelihood.

ForecastBench [15] is another dynamic benchmark designed to evaluate AI models' forecasting capabilities on real-world events whose outcomes are unknown at the time of prediction. Unlike traditional static benchmarks that rely on historical data and fixed test sets, ForecastBench constructs tasks from events such as elections, economic indicators, geopolitical developments, and other measurable phenomena, ensuring that the "test set" is effectively the future. Models are required to produce probabilistic forecasts, and evaluation is based on proper scoring rules that reward calibration and reliability, such as the Brier score or log-likelihood metrics.

CORE-Bench [6] is a benchmark designed to evaluate AI agents' ability to ensure computational reproducibility in scientific research. Unlike traditional benchmarks that focus on standard tasks, CORE-Bench constructs challenges based on real-world scientific research, requiring agents to replicate experimental processes and outcomes from published papers. Evaluation is based on criteria such as code execution, data handling, and result verification, ensuring that the reproduced results match the original findings. CORE-Bench aims to improve the credibility and transparency of scientific research.