Calibrated World Models for AI Agents: Prediction Market Data as Real-Time Context

Patrick Liu
SimpleFunctions
hello@simplefunctions.dev
https://simplefunctions.dev

April 2026

Abstract

Large language models have a knowledge cutoff that prevents them from reasoning accurately about current events. Existing mitigations—web search, news APIs, retrieval-augmented generation—return narrative text that requires parsing and provides no calibrated probabilities. We propose injecting prediction market data as a compact, structured world model into agent system prompts. Prediction markets aggregate the judgments of participants with real money at risk, producing calibrated probability estimates for geopolitical events, economic indicators, commodity prices, and elections. We introduce the **World Awareness Benchmark (WAB)**, a 44-question evaluation testing whether AI agents can accurately report current world conditions. Ground truth is derived from live prediction market prices. On WAB, a baseline Claude Haiku 4.5 model scores **2.3%**, while the same model augmented with an 800-token prediction market world state scores **70.5%**—a **31**× improvement. The world state injection requires no fine-tuning, no retrieval infrastructure, and adds only ∼800 tokens to the system prompt. We release the benchmark, daily world state snapshots, a Python SDK, and a public API as open resources.

1 Introduction

AI agents are increasingly deployed for tasks requiring awareness of current world conditions: portfolio analysis, geopolitical risk assessment, policy research, supply chain monitoring, and travel planning. These tasks share a common prerequisite: the agent must know, at minimum, the current state of geopolitical tensions, economic indicators, commodity prices, and political developments.

Current large language models (LLMs) cannot provide this reliably. Their training data has a cutoff date, and their outputs about current events are either hedged ("I don't have access to real-time data") or hallucinated—confidently stated but factually incorrect.

The standard mitigation is tool use, typically web search or news API integration. However, these sources have a fundamental limitation: they return *narrative text*, not *structured data*. Consider the difference:

- **Web search**: "According to recent reports, tensions in the Middle East remain elevated as diplomatic efforts continue..."
- **News API**: `{"title": "Iran tensions rise", "source": "Reuters"}`
- **Prediction market**: `Iran invasion: 53% (+5pp, $225K volume)`

The first two provide narrative that requires interpretation. The third provides a calibrated probability backed by real money—a number an agent can directly compare, threshold, and use in conditional reasoning.

We propose injecting prediction market data into agent system prompts as a structured world model. Our contributions are:

1. A **world state construction method** that selects anchor contracts by macro importance rather than price volatility, producing a compact ∼800-token representation of current world conditions.
2. A **delta update mechanism** that reduces ongoing context overhead from ∼800 tokens to ∼30–50 tokens per refresh cycle.
3. The **World Awareness Benchmark (WAB)**, a 44-question evaluation with live market ground truth, regenerated monthly.
4. **Empirical results** showing a 31× improvement in world awareness accuracy with no fine-tuning.

2 Background: Prediction Markets as Probability Sensors

A prediction market is an exchange where participants trade contracts on the outcomes of future events. A contract on "US recession in 2026" trading at 33 cents represents a 33% market-implied probability. The buyer pays 33 cents and receives $1 if the event occurs; the seller receives 33 cents and pays $1 if it does. Both have money at risk.

This mechanism produces well-calibrated probability estimates for several reasons:

Incentive alignment. Participants who overestimate probabilities buy contracts too expensive and lose money over time. Those who underestimate sell too cheaply. The market price converges toward the true probability because miscalibrated traders are systematically punished [Arrow et al., 2008, Wolfers and Zitzewitz, 2004].

Information aggregation. Prices aggregate private information from diverse participants—traders, analysts, domain experts, insiders—into a single number [Hayek, 1945]. This produces forecasts that outperform expert panels, surveys, and polls on average [Tetlock and Gardner, 2015, Surowiecki, 2004].

Continuous updating. Unlike surveys (weekly/monthly) or reports (quarterly), prediction market prices update in real time as new information arrives. A military escalation at 2 AM is reflected in prices within minutes.

Two major prediction exchanges operate at scale: **Kalshi** (CFTC-regulated, US-based, 5,000+ contracts) and **Polymarket** (blockchain-based, global, 4,700+ contracts). Together they cover geopolitics, economics, energy, elections, technology, cryptocurrency, and climate.

3 Method

3.1 World State Construction

We aggregate data from 9,706 prediction market contracts across Kalshi and Polymarket. Contracts are organized into six topic categories: Geopolitics, Economy, Energy, Elections, Crypto, and Tech.

A naive approach—selecting contracts with the largest 24-hour price changes—produces poor results. Daily close contracts ("Will natural gas close above $2.720 today?") have high volatility but near-zero information value for world state awareness. Conversely, critical macro contracts ("US recession probability") may move only 1–2 cents per day but represent the most important world state facts.

We address this with a two-tier selection approach:

Anchor contracts are selected by a composite importance score:

$$\text{score}(c) = \text{volume}(c) \times \text{macroBoost}(c) \tag{1}$$

where $\text{macroBoost}(c)$ assigns a $5\times$ multiplier to contracts matching fundamental macro keywords (recession, invasion, rate cut, war, GDP, nuclear, barrel) and a $0.1\times$ penalty to daily close contracts. This ensures that high-importance contracts always appear in the world state, regardless of daily price movement.

Mover contracts fill remaining slots per topic, ranked by volume$\times|\Delta\text{price}|$. Title-based deduplication prevents near-identical contracts (e.g., natural gas at \$2.720, \$2.725, \$2.745) from consuming multiple slots. A noise filter excludes sports, esports, weather, and social media metrics.

Output format. The resulting world state is $\sim$800 tokens of structured markdown, comprising:

- **SF Index**: Uncertainty (0–100, spread-based), Geopolitical Risk (0–100, geo market velocity), Momentum (-1 to $+1$, directional bias)
- **Traditional markets**: SPY, VIX, Gold, Oil, Treasury prices with daily change
- **Topic sections**: 2–4 contracts per topic with current price, 24h delta, and venue
- **Mispriced edges**: Top divergences between thesis model prices and market prices
- **Divergence alerts**: Cross-market anomalies (e.g., stocks and gold both rising)

3.2 Integration Patterns

We evaluate three integration patterns of increasing depth:

System prompt injection. The world state markdown is prepended to the system prompt. Cost: $\sim$800 additional input tokens. No tool calls required. This is the pattern evaluated in our benchmark.

Tool use with focused queries. The agent is given tools to request deeper coverage on specific topics. The `?focus=energy,geopolitics` parameter concentrates the same $\sim$800 token budget on fewer topics, yielding 10 contracts per topic instead of 4. A market search tool enables lookup of specific contracts.

MCP server integration. The world state is exposed via the Model Context Protocol [Anthropic, 2024], enabling zero-configuration discovery by compatible agent frameworks.

3.3 Delta Updates for Long-Running Agents

For agents operating in long sessions, re-reading 800 tokens of world state at each reasoning step is wasteful. We introduce a delta endpoint that returns only what changed since a given timestamp:

$$\text{delta}(t) = \text{worldState}(t_{\text{now}}) \ominus \text{worldState}(t) \tag{2}$$

The delta output is typically 30–50 tokens, covering index changes, significant market moves, new/resolved divergence alerts, and new contract entries. This provides a 16–20$\times$ reduction in per-cycle context overhead.

4 World Awareness Benchmark

4.1 Design

We introduce the World Awareness Benchmark (WAB), designed to test whether an AI agent can accurately report current world conditions. The benchmark consists of 44 questions across five categories (Table 1).

Table 1: WAB question categories with examples.

Category	N	Example Question
Geopolitical	10	What is the SF Geopolitical Risk Index?
Economy	10	What is the recession probability for 2026?
Energy	7	What is the current oil ETF (USO) price?
Elections	5	What probability for [candidate]?
Markets	12	What is the gold ETF (GLD) price?
Total	**44**	

4.2 Ground Truth and Scoring

Ground truth is derived from live prediction market prices at the time of evaluation. Each answer is scored on a 3-point scale:

- **2 points**: Exact match within tolerance (± 5 for probabilities in cents, $\pm \$5$ for asset prices)
- **1 point**: Correct direction or within $3\times$ tolerance
- **0 points**: Wrong, hallucinated, or refused to answer

Maximum score: 88 (44×2). The benchmark is **regenerated monthly** from live market data, ensuring it tests current knowledge rather than memorized facts from training data.

5 Experiments and Results

5.1 Setup

We evaluate Claude Haiku 4.5 [Anthropic, 2025] under two conditions:

- **Baseline**: Standard system prompt with no world context
- **+ World State**: System prompt augmented with $\sim$800 tokens of prediction market world state

Both conditions use identical prompting: "Respond with ONLY the number or value. No explanation." Temperature is set to 0. Each question is asked independently (no multi-turn context).

5.2 Aggregate Results

Table 2: WAB results for Claude Haiku 4.5 (April 2, 2026).

Condition	Score	Accuracy	Exact	Partial	Wrong
Baseline	2/88	2.3%	1	0	43
+ World State	62/88	70.5%	31	0	13
Improvement		**31**$\times$			

The baseline model answers almost every question incorrectly—it either hallucinates a number or states that it cannot access real-time data. With world state injection, 31 of 44 questions are answered exactly correct within tolerance.

5.3 Per-Category Analysis

Table 3: Per-category WAB accuracy.

Category	Baseline	+ World State	Δ
Economy	0/20 (0%)	16/20 (80%)	+80pp
Elections	2/10 (20%)	8/10 (80%)	+60pp
Markets	0/24 (0%)	18/24 (75%)	+75pp
Energy	0/14 (0%)	10/14 (71%)	+71pp
Geopolitical	0/20 (0%)	10/20 (50%)	+50pp

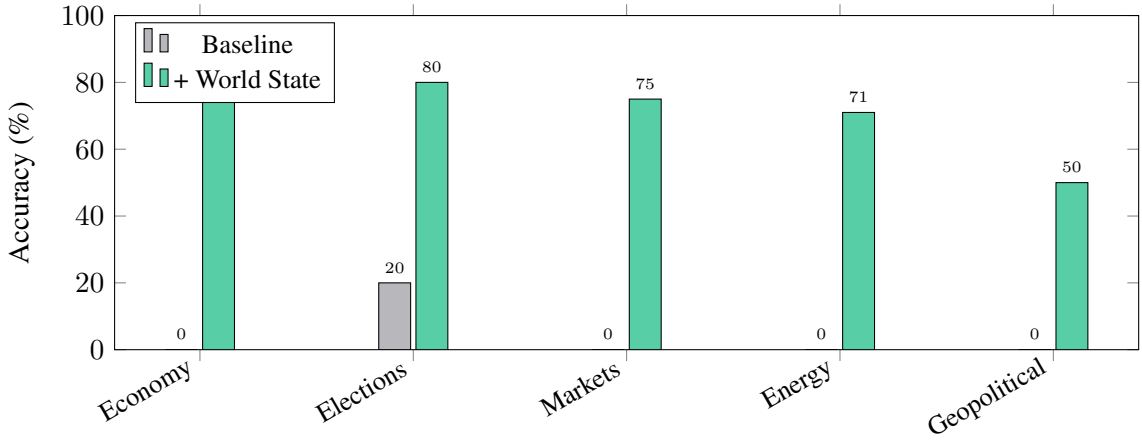

Figure 1: Per-category accuracy comparison. Economy and Elections show 80% accuracy with world state; Geopolitical is lower (50%) due to questions referencing specific contracts not present in the 800-token snapshot.

Economy and Elections show the strongest improvement (80%), likely because these categories have the most liquid and well-calibrated prediction market contracts. Geopolitical scores are lower (50%) because some benchmark questions reference specific contracts (e.g., a particular Iran sanctions contract) that are not among the top anchor contracts in the 800-token snapshot. This gap could be closed by using the tool-use integration pattern, where the agent can search for specific contracts.

5.4 Token Efficiency

Table 4: Comparison of context sources for agent world awareness.

Source	Tokens	Latency	Calibrated?	Structured?
Web search	2,000–5,000	2–5s	No	No
News API	500–1,000	500ms	No	Partially
RAG (recent docs)	1,000–3,000	1–3s	No	No
World state (full)	∼800	∼200ms	Yes	Yes
World state (delta)	∼30–50	∼100ms	Yes	Yes

The prediction market world state is 2.5–6× more token-efficient than web search and provides calibrated, structured data rather than narrative text requiring interpretation.

5.5 Error Analysis

The 13 incorrect answers in the world-state condition fall into two categories:

1. **Out-of-snapshot contracts** (9/13): Questions about specific contracts (e.g., a particular sanctions threshold) that were not among the anchor contracts in the 800-token snapshot. These would be answerable with the tool-use pattern.
2. **Parsing ambiguity** (4/13): The model returned a description instead of a number (e.g., "I cannot find this specific contract in the data"), which was scored as 0 despite being an honest acknowledgment of missing information.

No hallucinated answers were observed in the world-state condition—when the model lacked data, it said so rather than fabricating a number.

6 Related Work

Prediction market calibration. The calibration properties of prediction markets are well-established in the economics literature. Arrow et al. [2008] summarize the theoretical foundations. Wolfers and Zitzewitz [2004] provide empirical evidence of calibration across political and economic markets. Tetlock and Gardner [2015] show that prediction markets outperform expert forecasters on average. Our work applies these calibrated probabilities as context for LLM agents rather than as standalone forecasting instruments.

LLM grounding and tool use. Retrieval-augmented generation [Lewis et al., 2020] and tool use [Schick et al., 2023] address the knowledge cutoff problem but typically rely on unstructured text retrieval. WebGPT [Nakano et al., 2021] uses web browsing but returns narrative text. Our approach provides structured, calibrated data requiring no parsing or interpretation by the model.

Agent architectures. ReAct [Yao et al., 2023] and related work [Wang et al., 2023] focus on reasoning and tool-use patterns but do not address the world awareness gap. The world state is complementary to these architectures—it can be injected into any agent's system prompt as an additional context source.

Real-time data for LLMs. Recent commercial efforts (Perplexity, Google's grounding with Search) provide real-time web data but return narrative text, not calibrated probabilities. Our approach is the first we are aware of that uses prediction market prices as structured context for LLM agents.

7 Limitations

- **Market coverage**: Not all events have active prediction market contracts. Coverage is strongest for US geopolitics, economics, and elections. Regional events in developing economies may be underrepresented.
- **Thin markets**: Low-volume contracts may not be well-calibrated. Our anchor selection mechanism mitigates this by prioritizing high-volume contracts, but edge cases exist.
- **Manipulation risk**: Prediction markets can theoretically be manipulated, though the cost of sustained manipulation is high and detection mechanisms exist [Hanson, 2006].
- **Update frequency**: The world state updates every 15 minutes. Events that unfold faster (e.g., flash crashes, breaking military actions) may not be reflected immediately.

- **Benchmark scope**: WAB tests factual recall of current probabilities, not the agent's ability to *reason* over them. Future work should evaluate downstream task performance (e.g., portfolio rebalancing decisions conditioned on world state).
- **Single model**: We evaluate only Claude Haiku 4.5. Results may differ for other models, though we expect the direction of improvement to be consistent since the mechanism (context injection) is model-agnostic.

8 Conclusion

Prediction market data provides a compact, calibrated, and structured world model for AI agents. At ∼800 tokens, it is more token-efficient than web search, more calibrated than news, and more structured than RAG. The World Awareness Benchmark provides a reproducible way to measure the gap between what agents know and what is currently true—and to measure the effectiveness of context injection in closing that gap.

Our results demonstrate that the world awareness problem is not a model capability problem but a *context problem*. The same model that scores 2.3% without context scores 70.5% with 800 tokens of structured data. This suggests that investment in better world state construction—selecting, compressing, and structuring real-time data for LLM consumption—may be more impactful than scaling model parameters for current-events reasoning.

We release all resources as open infrastructure for the AI agent ecosystem.

Resources

- World State API: `https://simplefunctions.dev/api/agent/world` (free, no auth)
- Benchmark: `https://huggingface.co/datasets/SimpleFunctions/world-awareness-benc`
- Daily snapshots: `https://huggingface.co/datasets/SimpleFunctions/world-state-daily`
- Python SDK: `pip install simplefunctions-ai`
- MCP Server: `https://simplefunctions.dev/api/mcp/mcp`
- This paper: `https://simplefunctions.dev/papers/world-model`

References

Anthropic. Model Context Protocol specification, 2024. https://modelcontextprotocol.io.

Anthropic. Claude 4.5 model family, 2025. https://docs.anthropic.com/en/docs/about-claude/models.

Kenneth J Arrow, Robert Forsythe, Michael Gorham, Robert Hahn, Robin Hanson, John O Ledyard, Saul Levmore, Robert Litan, Paul Milgrom, Forrest D Nelson, et al. The promise of prediction markets. *Science*, 320(5878):877–878, 2008.

Robin Hanson. Designing real terrorism futures. *Public Choice*, 128(1):257–274, 2006.

Friedrich A Hayek. The use of knowledge in society. *American Economic Review*, 35(4):519–530, 1945.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *NeurIPS*, 2020.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *NeurIPS*, 2023.

James Surowiecki. *The Wisdom of Crowds*. Doubleday, 2004.

Philip E Tetlock and Dan Gardner. *Superforecasting: The Art and Science of Prediction*. Crown, 2015.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *arXiv preprint arXiv:2308.11432*, 2023.

Justin Wolfers and Eric Zitzewitz. Prediction markets. *Journal of Economic Perspectives*, 18(2):107–126, 2004.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. *ICLR*, 2023.