Reasoning and Tools for Human-Level Forecasting

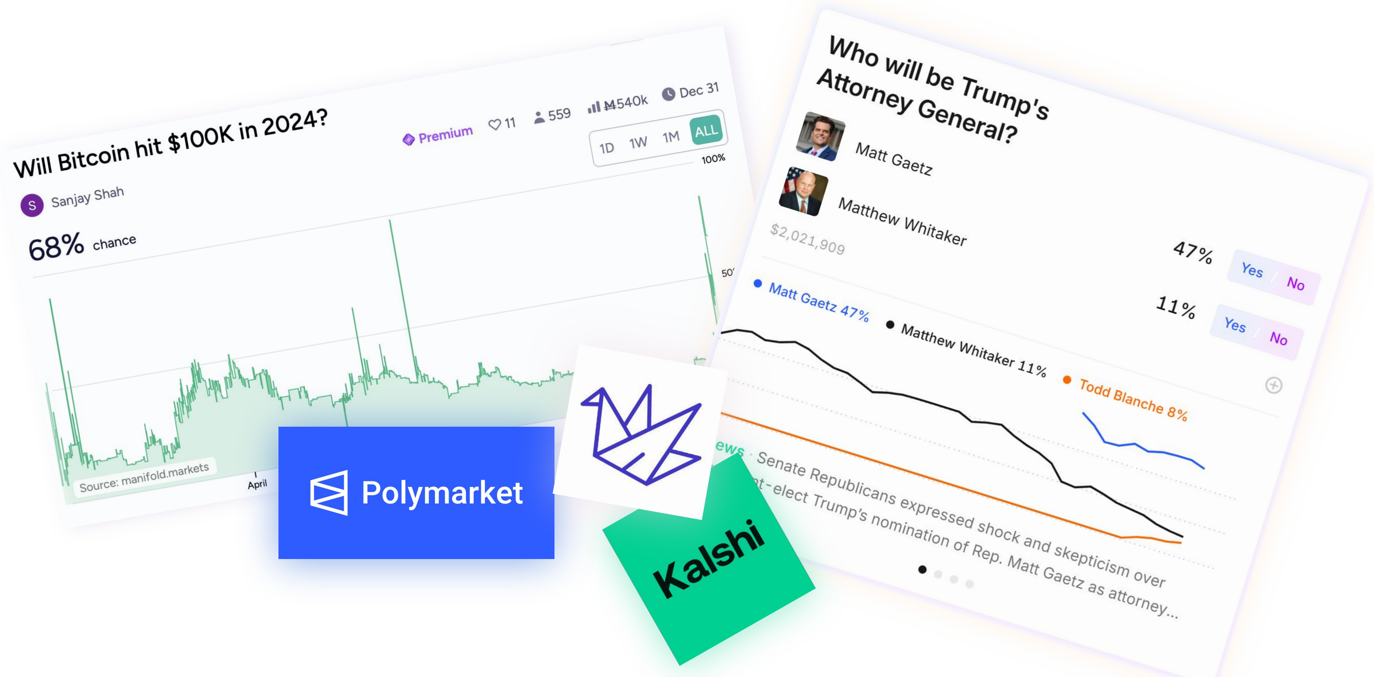
Elvis Hsieh*, Preston Fu*, Jonathan Chen* (equal contribution)
UC Berkeley

Why LLM forecasting?

Large language models (LLMs) trained on web-scale datasets **are**:
- Good at memorizing large amounts of training data, even if only present in a few examples.
- Often evaluated on tasks such as question-answering, which demonstrate world knowledge but not reasoning capabilities.

With RTF, LLMs **can be**:
- Good at reasoning in live settings, when presented with real-time data and a basis for truth.
- Successful in difficult, reasoning-intensive decision-making tasks like forecasting.

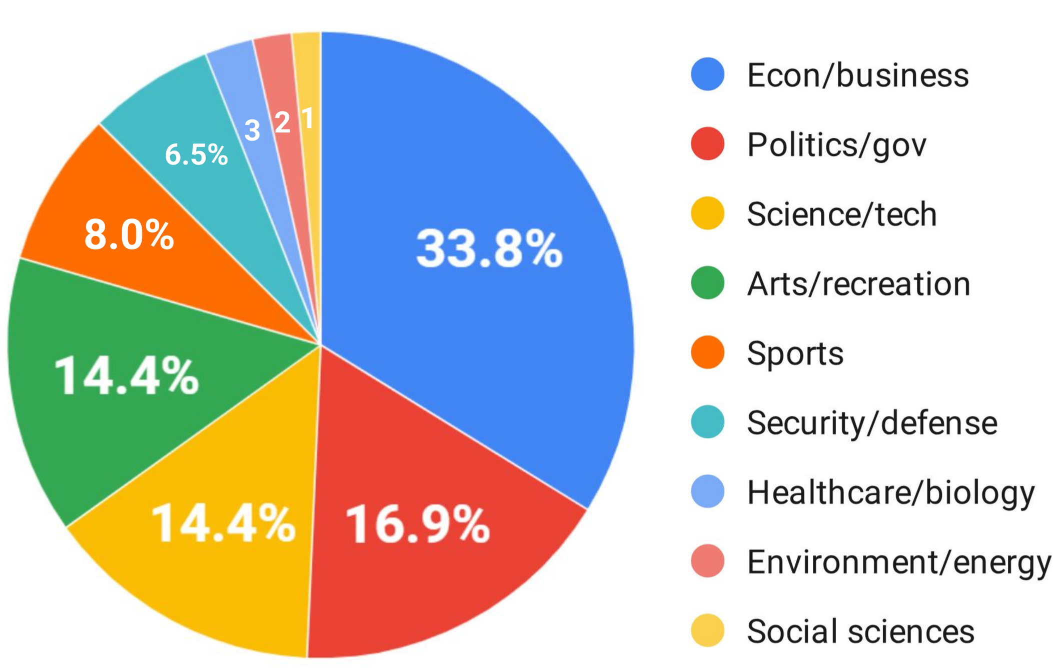
We propose a **zero-shot prompting** mechanism achieving **human-level forecasting performance**.



Dataset

201 questions from Manifold Markets.

Sample question: "Will ETH close above $3700 on April 30, 2024?"



Experiments

Method	Brier ↓	Acc % ↑	Std ↓
Crowd	0.172	73.8	
RTF Median of 3	**0.169**	72.4	0.092
RTF Mean of 3	0.170	**73.9**	0.092
RTF Sampled	0.180	71.6	
Halawi et al. (2024) GPT-4o	0.177	68.7	
GPT-4o	0.210	65.5	
Base LM Mean	0.218	62.9	0.150
Base LM Median	0.228	61.3	0.150
Llama 3	0.256	56.2	
GPT-3.5	0.261	53.5	
GPT-4	0.265	54.8	

Brier score: $BS = \frac{1}{n}\sum_{i=1}^{n}(f_i - o_i)^2$ (how accurate are forecasts?)

Method	Calibration Index ↓
Crowd	0.0101
ReAct Mean	**0.0129**
ReAct Median	0.0137
ReAct	0.0164
GPT-4o	0.0194
GPT-4	0.0290
GPT-3.5	0.0298
Llama 3	0.0301

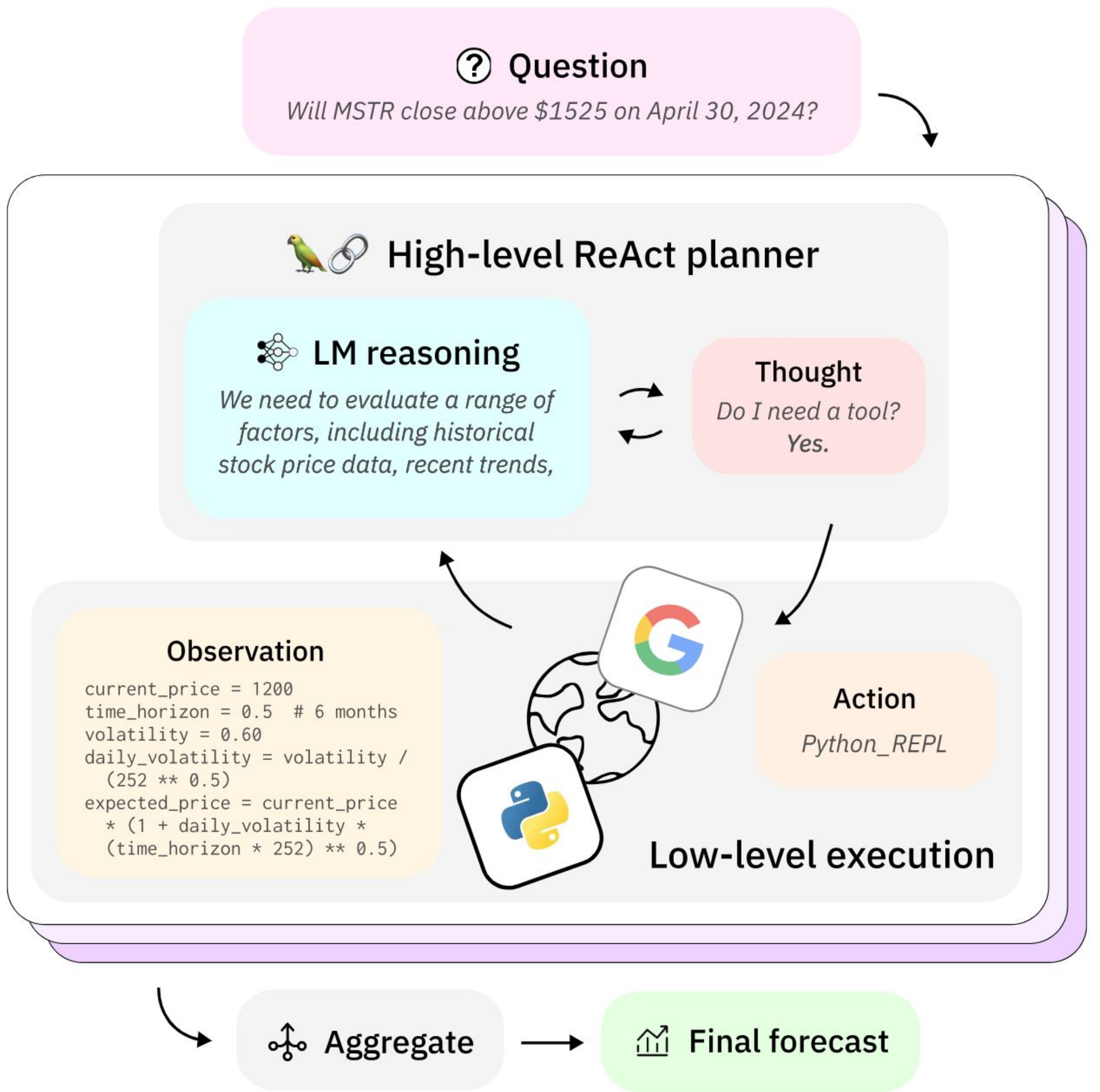
Calibration index: $CI = \frac{1}{N}\sum_{k=1}^{K} N_k(f_k - o_k)^2$ (how close are predictions to binned outcome frequencies?)

Background

- Naively prompting LLMs for forecasting tasks performs worse than humans (prediction markets are good data source). [1]
- Following the **wisdom of crowds effect** of humans, large aggregates (size up to 36) of LLM predictions work better than individual LLMs. [2]
- **Reasoning-and-acting** (ReAct), unlike chain-of-thought, continuously refines responses with retrieved information. [3]

We show that ReAct-based frameworks are suitable for forecasting tasks.

Method



A **small ensemble** of **hierarchical agents**:
- **High-level agents** act as planners, handling abstract logic and forecasting principles to aggregate information.
- **Low-level agents** generate inputs to tools (Google, Python), execute the actions, and report observations.
- Delegating reasoning and API calling to specialized agents enhances efficiency, conserves tokens, and allows for more complex operations.
- RTF is simple and scalable, and can achieve good performance on different data and LLMs. No need for fine-tuning!

Analysis

- Small ensembles of highly accurate agents are sufficiently good. **One RTF agent is better than an aggregate of low-accuracy agents!**
- Base LLMs produce higher-variance outputs compared to RTF. Ensemble performance is limited by base LLM reasoning.
- Qualitative assessment: direct prompting produces cascading errors (most recent tokens matter more), while RTF yields more **cohesive, human-like reasoning trajectories**.

References

[1] A. Zou, et al. Forecasting Future World Events with Neural Networks. Preprint, arXiv:2206.15474.
[2] P. Schoenegger, et al. Wisdom of the silicon crowd: Llm ensemble prediction capabilities rival human crowd accuracy, 2024.
[3] S. Yao, et al. React: Synergizing reasoning and acting in language models

Paper