

Prediction Markets

Tested against people betting real money

How FishDog checks its synthetic population against real-money markets, and what the benchmark is and is not.

CONTENTS

Why prediction markets

One market, start to finish

A reliability profile before the event settles

What this benchmark is, and what it is not

A survey records what people say. A prediction market records what people are willing to bet on, and every market ends in a fact. FishDog puts the same questions traders are betting on to its synthetic population, and keeps score.

Why prediction markets

Real money is the counterparty

Traders lose money when they are wrong, so a market price carries information a survey answer does not. It is a demanding benchmark to sit next to.

Every market settles

The event happens or it does not. There is no arguing about methodology after the fact, which makes it a signal-versus-outcome test.

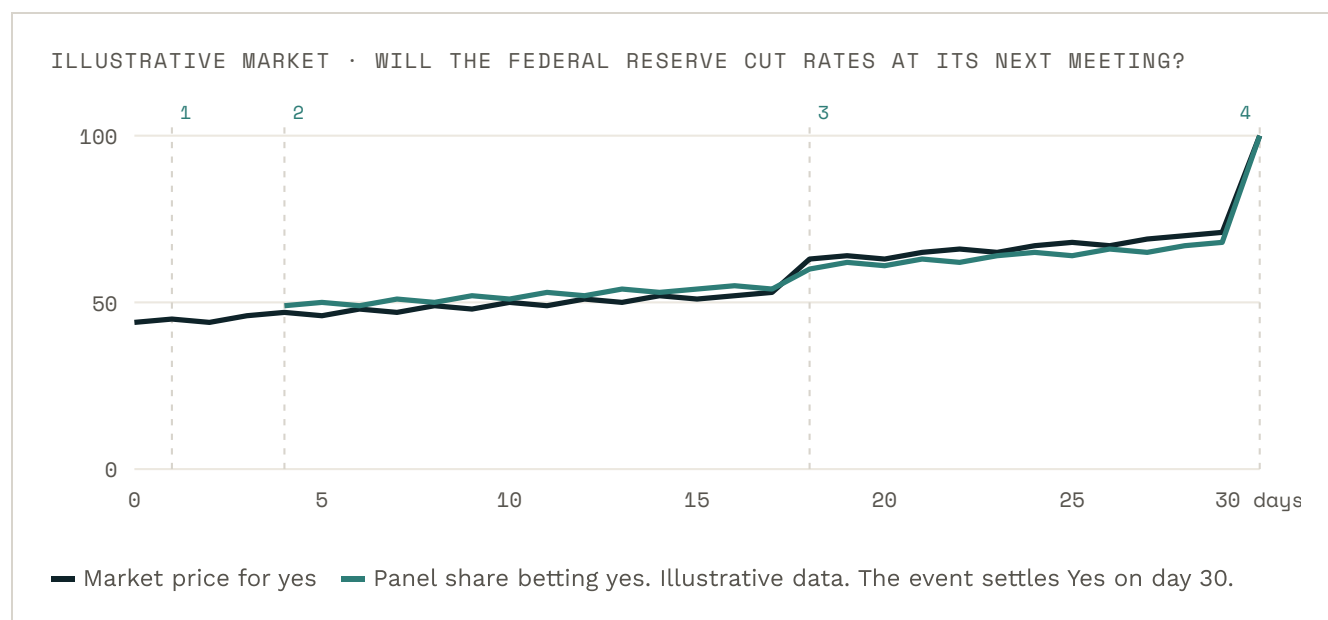
The answer is not in the past

The tests run on events that postdate the model's training, so the population cannot recall the answer. It has to reason its way there.

Context. Real-money markets have a long record as forecasters: the Iowa Electronic Markets, an academic market, beat contemporaneous polls 74 percent of the time across 964 polls (Berg, Nelson, and Rietz, 2008). FishDog uses that literature as context, not as its own evidence.

One market, start to finish

The life of one market, and how it becomes a test. The numbers mark the four moments below.



1. Day 1 - A market worth testing against

FishDog imports live events from Polymarket. Suitability rules filter out thinly traded markets and keep questions that apply to Americans broadly. This one is a real-money question about an event that has not happened yet.

2. Day 4 - The same question goes to the panel

A panel of 100 Americans, recruited from FishDog's digital twin of the US population, gets the question the market is trading on. Each synthetic person places a bet in FishDog's own database. Nothing is placed on Polymarket itself.

3. Day 18 - News breaks, and both reprice

Traders move the price within hours. The panel reads its own news, so it can be re-run the same day and its answer moves too. The test is whether the two move together, and by how much.

4. Day 30 - The event settles, and the score is filed

The outcome is a fact neither side can argue with. The panel is scored, including Brier score and calibration error, and the question is tagged on 18 axes. The score joins every earlier market of the same shape on the map.

A reliability profile before the event settles

Because every market is scored on the same 18 axes, a new question can be compared with earlier questions of the same shape. That gives a reliability profile while there is still time to act.

QUESTION	AXES	PROFILE	WHY
Will an energy company hit its production target this quarter?	Business, Technical, Near future, Macro	Stronger	Technical and expertise-driven events are where the population is strongest. Questions like this one resemble many earlier markets that settled.
Will the Federal Reserve cut rates at its next meeting?	Economics, Macro, Near future, High stakes	Stronger	Economic data is a strong area. The paper lists it among the areas the feedback loop leaves alone.
Which party wins a closely watched governor's race?	Politics, Macro, Near future, High stakes	Stronger	Politics is a strong area, and this question shares its shape with many earlier political markets.

QUESTION	AXES	PROFILE	WHY
Will a new phone model ship before the holidays?	Science and technology, Product, Near future, Micro	Stronger	Product questions are one of the paper's examples of where the population is strongest, alongside energy.
Will a small-town council approve a new parking rule this month?	Politics, Micro, Near future, Low information	Weaker	Low-information events, where there is little public signal to reason from, are the weakest area. The profile says so before anyone relies on the answer.

What this benchmark is, and what it is not

A benchmark, not training data

Prediction markets are what the population is checked against. The markets themselves are not what it is built from. The population comes from census data and its own news sources.

A pattern, not a single call

One market is one event, and a single event cannot be back-tested. The value is in the accumulation: many markets, scored on the same axes, showing where the population holds and where it does not.

One layer of four

Prediction markets sit alongside the census construction floor, Gallup and Pew comparisons, and the University of Michigan consumer-sentiment series. The Synthetic Populations paper sets out all four.

Weak spots stated up front

Low-information markets, where there is little public signal to reason from, are the weakest area, and FishDog says so. When the map shows a gap, FishDog adds signal and the gap is tracked until it closes. A profile that admits its weak spots is one a reviewer can use.

References

1. Berg, J. E., Nelson, F. D., & Rietz, T. A. (2008). Prediction market accuracy in the long run. *International Journal of Forecasting*, 24(2), 285–300.
2. Gales, P., & Duess, A. (2026). [Synthetic populations: How FishDog builds, calibrates, and validates them.](#)

The interactive version of this explainer is at fish.dog/validation/prediction-markets.