Synthetic Experts

How Thesis Lab Builds, Calibrates, and Validates a Recruited Population

A FishDog methodology and validation paper
For institutional and professional investors. Not investment advice.

Authors: Phillip Gales (Chief Product Officer) and Andreas Duess (CEO)
June 2026

Contents

Executive summary

The primary-research inputs hedge funds rely on are getting noisier, slower, and more expensive. Analysts tell us the decline is already visible in the workflow: expert-network rosters converge across funds, some paid calls now produce ChatGPT-assisted answers from people who are not the experts they claim to be, and cheap panels are increasingly hard to trust.

Human research remains valuable, but the "real human" baseline is no longer a clean benchmark. Survey panels self-select toward respondents who value their time at a low dollar amount; recruitment fraud and coached personas reportedly add more noise; and expert calls remain useful but uneven, expensive, and difficult to scale. The buyer's real question is not whether synthetic research can replace human judgment. It is whether a synthetic method can deliver a fast, inspectable signal with known failure modes.

Thesis Lab in one page.

What it is	A synthetic expert network for investment research. An analyst poses the question they would take to an expert call; Thesis Lab recruits relevant synthetic experts from FishDog's calibrated population, interrogates them at scale, and returns the answer with the reasoning behind it.
What it is not	Not a generic LLM or chatbot, not an "LLM wrapper," not a source of material non-public information, and not a replacement for analyst judgment. It holds no private, single-company facts by construction.
Why it matters	The primary-research inputs funds rely on are degrading: expert-network rosters converge across funds, some paid calls now return chatbot-assisted answers, and cheap panels are harder to trust, all while fielding costs rise and the half-life of a signal shrinks.
How it fits the workflow	A leading indicator and complement, not a substitute. Before an expert call (shape and pressure-test the hypothesis), between prints (a fast cohort read), on any micro-cohort on demand, and after a market-moving event (re-run the same panel the same day).
How value is measured	Speed to signal, cost versus expert calls and panels, quality of hypothesis generation, repeatability of a pinned re-run, breadth of evidence, and usefulness in investment-committee work.
How it is trusted	A calibrated population benchmarked against external yardsticks (Census and ACS, real-money prediction markets, Gallup and Pew, and University of Michigan sentiment), with version control, disclosed panel composition, and stated limitations on every study.

In this paper:

- **FishDog** is the company and methodology layer: it builds and validates calibrated synthetic populations for a wide variety of applications.
- **Thesis Lab** is FishDog's investment-research product: a synthetic expert network built on that population infrastructure.
- **Eagle Alpha** is the launch partner for institutional alternative-data buyers; FishDog remains responsible for the methodology and validation claims in this paper.

FishDog builds a statistically valid representation of a population, calibrated from the most recent United States Census and ACS Public Use Microdata and enriched against 40+ additional data sources. Each expert carries role, location, current-news, and local-condition inputs, plus OCEAN five-factor personality traits that shape how they interpret information and respond.

Those experts can be interrogated like an expert call, re-cut into cohorts on demand, re-run when the news changes, and benchmarked against external yardsticks including real-money prediction markets, Gallup and Pew, and the University of Michigan consumer-sentiment series.

Methodology at a glance

Population frame	US Census and American Community Survey (ACS/PUMS) microdata, calibrated to ~1% average marginal error on selected marginals; enriched with 40+ role, location, news, local-condition, and personality inputs.
Currency / lag	Experts reason from a live feed of 359 news and data sources, reflecting new information within roughly four hours of its arrival rather than from a static base-model snapshot.
Validation	Layered and continuous: prediction markets across 18 event axes (daily), Gallup and Pew (~9% marginal error), and the University of Michigan consumer-sentiment series.
Cadence	Cohorts recruited and fielded on demand; benchmarks scored daily; a panel can be re-run the same day after a market-moving event.
Disclosure standard	Every study reports panel composition, observed versus inferred attributes, run date, and model, prompt, population, and source versions.

Key points:

- **Recruit, do not create.** Drawing from a calibrated population reduces a major source of hand-selection bias. The validation stack then measures the model, construction, and prompt biases that remain.
- **A layered validation stack rather than a single accuracy number.** The population is built from Census and ACS microdata, enriched with 40+ role, location, current-news, local-condition, and personality inputs, and then benchmarked against real-money prediction markets, Gallup and Pew survey data, and the University of Michigan consumer-sentiment series. The output created is based on the combined evidence.
- **A reliability estimate before the event resolves.** Because every benchmark question is scored on the same 18 prediction-market axes, a new question can be matched to prior questions of similar shape, giving an expected reliability range before the output is treated as signal. The analyst can read how much to trust an answer while there is still time to act, not after the event settles.
- **Known limitations.** Synthetic respondents are documented in the literature to flatten variance, to be sensitive to prompt wording, and to drift across model versions. We address each directly and describe what we do about it.
- **Speed, cost, and re-cuttability.** Any cohort, fielded today, interrogated like an expert, and re-run after a market-moving event, at a fraction of the cost and time of a human panel.

This paper sets out how the population is built, how experts are recruited from it, how those experts are calibrated and given role-specific context, how outputs are tested, and where the method is weak. The intended readers are investors, data-science reviewers, and compliance teams who will inspect the assumptions.

---

1. The problem: primary-research inputs are degrading

Most conversations about synthetic research start from a faulty premise: that human research is ground truth and everything else must be graded against it. That premise no longer holds.

Expert networks are converging and degrading. The economics of the expert-network model push every fund toward the same roster. When three to six funds interview the same expert about the same question, the resulting "edge" converges and stops being edge. Worse, the screening problem has changed. Non-experts can, and reportedly do, join a paid call with a chatbot running and feed back AI-generated answers, funds we interviewed for this paper increasingly treat this as a diligence problem.

Survey panels are contaminated. Paid online panels self-select toward respondents who place a low value on their time, which skews the sample away from the population an investor cares about.

Some recruitment agencies reportedly coach participants to play-act the personas required to qualify for paid studies, so a meaningful share of the "real humans" in a panel may not even be who they claim to be. Add the well-documented problems of social-desirability bias, acquiescence, and panel conditioning, and the baseline looks less like ground truth and more like a noisy, biased instrument that happens to be made of people.

Both are slow and expensive, at exactly the moment fielding costs are rising and the half-life of a tradable signal is shrinking.

Properly governed human research remains valuable, but the comparison a buyer wants to make, synthetic versus "real," is against a baseline that is itself broken in very specific ways which will continue to degrade results.

A synthetic population built by construction does not self-select, cannot be coached to defraud the screener, and does not try to tell the interviewer what it thinks the interviewer wants to hear. It has its own failure modes, addressed in Section 5, but they are different failure modes which can be measured and corrected.

Exhibit 1. Where the two incumbent primary-research inputs break down.

Failure mode	Expert networks	Survey panels
Roster / sample formation	The same experts are reused across funds, so the edge converges as more buyers hear the same view.	Respondents self-select into paid panels, often because their time is cheap enough to make low-paid participation worthwhile.
Screening integrity	Non-experts can qualify for calls and use AI tools during interviews, making expertise harder to verify.	Recruitment fraud and coached personas make it harder to know whether respondents are who they claim to be.
Signal quality	A strong call can be deep, but the output depends on one person's recall, incentives, and willingness to speak plainly.	Panels provide breadth, but responses are shallow, self-reported, and exposed to social-desirability bias, acquiescence, and panel conditioning.
Speed and cost	Scheduling, screening, compliance review, and premium hourly rates create lag and cost.	Lower per respondent cost, but quality control, re-fielding, and fraud management add friction.
Follow-up / re-cutting	Follow-up is rich but requires more calls, more scheduling, and more budget.	Cohorts are fixed once fielded; new cuts or follow-up questions often require another wave.
Compliance / provenance	MNPI exposure is a standing risk because the value comes from human insiders and operators.	MNPI risk is lower, but respondent provenance and identity remain difficult to verify at scale.

Source: FishDog analysis. Failure modes drawn from the survey-methodology literature (see Section 5 and References) and from interviews with funds conducted for this paper.

---

2. What Thesis Lab is

FishDog is the company and methodology layer: it builds, enriches, and validates calibrated synthetic populations. Thesis Lab is the investment-research product built on top of that infrastructure, launched in partnership with Eagle Alpha for institutional alternative-data buyers.

An analyst poses the kind of question they would normally take to an expert call, and Thesis Lab recruits relevant synthetic experts from FishDog's calibrated population, interviews them at scale, and returns both the answer and the reasoning behind it.

Recruit, do not create. The common way to build synthetic respondents is to write personas: describe a few archetypes and prompt a model to speak as them. That approach imports two biases at once, the author's idea of who the customer is and the model's tendency to tell a tidy story.

FishDog inverts the order. We build a statistically valid representation of an entire population first, calibrated to real demographic distributions, and then recruit a panel from it the way a research firm would recruit from the real world.

Recruitment from a calibrated population still leaves some inherent bias in the system, based on training data. But what it removes is the largest single source, hand-selection bias, and it makes what remains auditable. Every expert is the output of a method that starts from a defensible population frame and gets tested continuously against external benchmarks.

Because recruitment starts from a population rather than a hand-written persona, Thesis Lab can surface the kinds of experts a fund often finds hardest to reach: the warehouse manager who plausibly runs cold-chain logistics for a national restaurant account, or the oncology nurse coordinating infusion care in a regional cancer clinic. Hard-to-reach expertise is often where the human expert-network model is thinnest and most expensive, and where a recruited synthetic population has the most to add.

The prediction-market benchmarks in Section 4 demonstrate the same outcomes, with technical and expertise-driven questions scoring better than low-information consumer or gossip questions. For analysts, Thesis Lab is most useful when the expertise barrier is high.

Three properties follow from the design and distinguish it from a human panel:

- **On-demand cohorts.** A fixed survey is fixed at fielding. Thesis Lab can generate and probe cohorts such as chief information officers at firms of 250 to 500 employees in a named vertical, without recruiting them one by one.
- **Interrogation over one-shot capture.** An expert in the panel can be asked follow-up and "why" questions individually and at volume, surfacing the reasoning behind a number. A fixed-form survey cannot do this without re-fielding.

- **No fatigue, no attrition, no recruitment lag.** Repeated waves do not burn the panel. Consistency over time is controlled through software, versioning, and validation rather than recruited for each wave.

Additionally, there is a structural compliance advantage that matters specifically to this use case. A synthetic expert does not possess material non-public information about any company, because it has none to possess. The human expert-network model carries the reverse risk by design, which is why funds staff whole compliance teams around it. Thesis Lab materially reduces that exposure by construction, while also setting a hard boundary: Thesis Lab is not a source of private, single-company facts precisely because it has none.

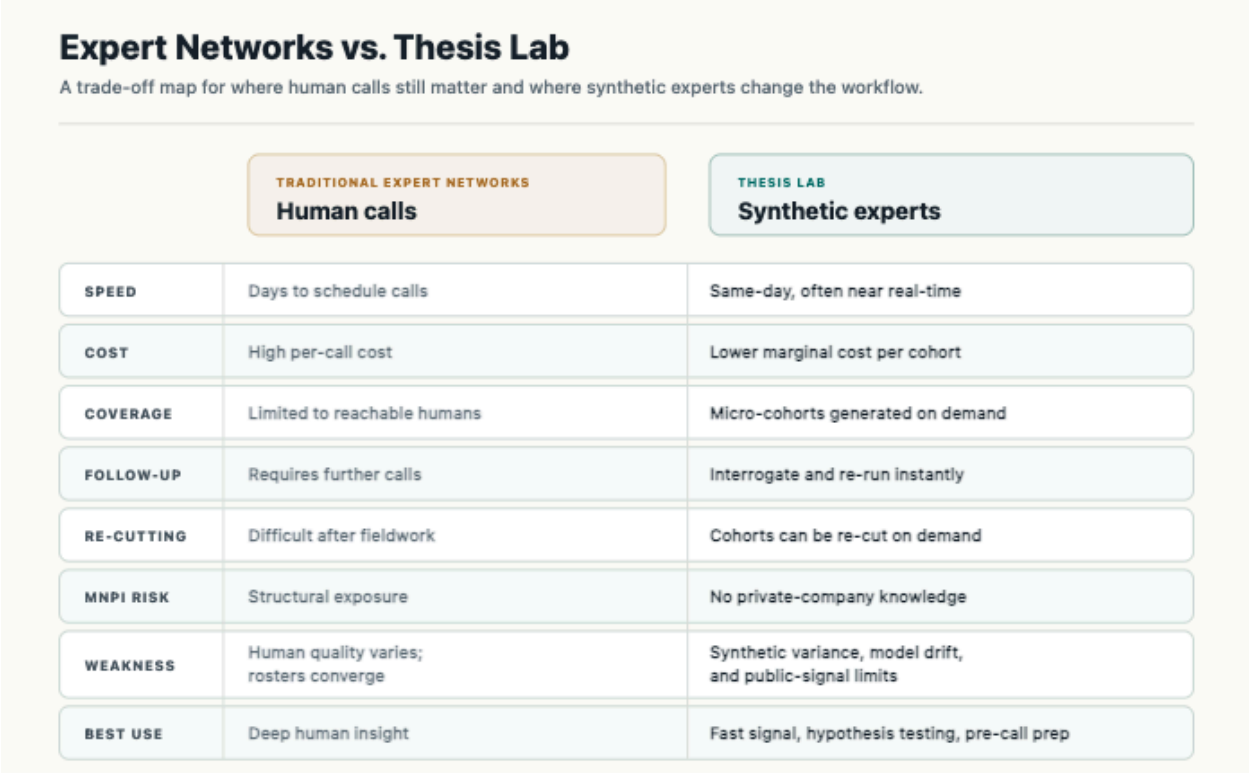
Thesis Lab sits alongside the incumbents, a faster and cheaper complement and leading indicator to human research, with its credibility resting on the validation in Section 4.

Exhibit 2. Create-a-persona versus recruit-from-a-population.

Create a persona	Recruit from a population
Start with an archetype the author imagines: "a CIO," "a soccer dad," "a rural patient."	Start with a calibrated population built from Census, ACS/PUMS, and additional role, location, news, and personality inputs.
The analyst or prompt writer decides who exists before the research begins.	The population defines who can be recruited; the analyst specifies the research question and eligibility criteria.
The model is asked to role-play the archetype.	Thesis Lab recruits matching experts from the population complete with role-specific public context.
Bias enters through persona design and the model's tendency to produce a tidy, average answer.	Bias is constrained by the population frame, source versioning, and external validation. Remaining bias is measured rather than assumed away.
Outputs are difficult to audit because the respondent was invented for the prompt.	Outputs can be traced to a population frame, cohort definition, source bundle, model version, and benchmark history.

Source: FishDog methodology.

Exhibit 3. Traditional expert networks versus Thesis Lab synthetic experts.



Expert Networks vs. Thesis Lab

A trade-off map for where human calls still matter and where synthetic experts change the workflow.

	TRADITIONAL EXPERT NETWORKS **Human calls**	THESIS LAB **Synthetic experts**
SPEED	Days to schedule calls	Same-day, often near real-time
COST	High per-call cost	Lower marginal cost per cohort
COVERAGE	Limited to reachable humans	Micro-cohorts generated on demand
FOLLOW-UP	Requires further calls	Interrogate and re-run instantly
RE-CUTTING	Difficult after fieldwork	Cohorts can be re-cut on demand
MNPI RISK	Structural exposure	No private-company knowledge
WEAKNESS	Human quality varies; rosters converge	Synthetic variance, model drift, and public-signal limits
BEST USE	Deep human insight	Fast signal, hypothesis testing, pre-call prep

Source: FishDog analysis. This is a trade-off map, not a claim of superiority: Thesis Lab's weaknesses are detailed in Section 5, and the two methods are complements (Section 7).

---

3. How the experts are built and calibrated

The build follows a fixed order: construct the population, recruit and enrich the experts, add role-specific context, give them texture and memory, and keep them current.

Population construction

The population is calibrated against the most recent United States Census (for US personas) and the American Community Survey Public Use Microdata Sample, which provides individual record-level microdata across occupation, health, income, and dozens of other attributes. We build calibration targets from PUMS and align the synthetic population to them, achieving a distributional fit of roughly 1% marginal error on average against selected census-frame marginals.

The method is standard survey statistics. Post-stratification and raking to known population marginals are what national pollsters use, both well established in the literature (Little, 1993; Pew Research Center, 2018). The American Community Survey reports margins at the 90% confidence level, and a national poll at n=1,000 runs to roughly plus or minus three points.

A note on scope. This paper describes the United States population, because the public ground-truth instruments used in validation, the Census, Gallup and Pew, and the University of Michigan series, are richest and longest-running in the US, which makes it the most demanding place to show the method.

The same construction approach is already being applied to other markets, calibrated to each market's own authoritative demographic sources rather than to US benchmarks. Validation in each market is anchored to the instruments available there, so the depth of external benchmarking varies by how much public ground truth a given country publishes.

From population member to recruited expert

A census-calibrated population member becomes a useful expert by separating four different things: what is directly grounded in data, what is modeled from that data, what public information the expert is allowed to use, and how it is required to answer.

1. **Observed and calibrated attributes.** These are attributes grounded directly in the census and ACS/PUMS frame, such as age, geography, education, household structure, income band, occupation class, and selected health and labor-market variables.
2. **Inferred and enriched attributes.** These are attributes that are plausible conditional on the observed frame but not directly present in the census microdata, such as role seniority, budget authority, firm-size band, procurement exposure, or whether a worker in a given occupation is likely to interact with a specific category of vendor. These attributes are generated under constraints and audited for distributional reasonableness. They

should be treated as modeled enrichment, but not observed fact taken from specific human reference points.

3. **Role-specific source context.** Once an expert is recruited, the system gives that expert access to the public information their real-world counterpart would plausibly consume: trade publications, public filings, relevant news, regional conditions, domain literature, and other sources. This supplies context for the expert's reasoning.
4. **Question-time constraints.** At answer time, the expert is instructed to distinguish what it knows from public information, what it infers from role experience, and what it cannot know. That constraint matters most for company-specific questions, where Thesis Lab can provide structured inference but not private disclosure.

Each layer does one specific job:

- **Census calibration** sets the population frame.
- **Enrichment** makes that frame useful for recruiting experts.
- **Role-specific source context** gives a recruited expert current information to reason from.
- **Validation** checks whether the resulting signal tracks the outside world.

Recruitment and role-specific context

Every expert shares a common baseline: demographics flow to occupation, income, and skills, which flow to health, which flows to the information that person would plausibly consume. Specialists then receive role-specific public material. An AI professor and an oncology nurse start from the same construction and diverge through the source feeds each is given. No expert is inherently "more powerful" than another; the difference is the role-specific context each has available.

We keep this source grounding separate from fine-tuning. Source grounding adds or updates public materials and is reversible. Fine-tuning changes the model weights, and we describe it as such when we use it.

Texture and personality

Experts are given texture so their outputs read as messily human rather than uniform. Each consumes role-appropriate and location-appropriate news and weather; each carries a personality grounded in the five-factor (OCEAN) model, with trait distributions drawn from real data. The five-factor model is the field's consensus descriptive taxonomy of personality (John & Srivastava, 1999; Goldberg, 1990; McCrae & Costa, 1987).

We extend that validated five-dimension foundation with additional modelled dimensions; we are explicit that the additional dimensionality is a FishDog design choice and that the academic literature validates the five-factor foundation.

Memory

Memory is separated into two layers to prevent client interactions from contaminating the product.

Persistent memory holds the expert's stable profile and external context: character, macro conditions, news, weather, and health. It changes only when outside conditions change. An oil-price shock can affect an expert in an energy-exposed role; a client's line of questioning cannot.

Episodic memory is limited to a single session and is cleared when that session ends. Client interactions are logged for the client's own use, but they are not fed back into the global product.

Staying current

Base language models are trained on a static snapshot of the world, and the lag varies by model and vendor. Thesis Lab closes that gap with a live feed of 359 news and data sources, so experts reflect new information within roughly four hours of its arrival rather than reasoning from a stale base-model snapshot. For specialist briefs, the feed can inject domain literature, for example research articles for a clinical panel, in place of general news.

The feed is curated by profile. Source bundles vary by role, region, demographics, and worldview, reflecting observed media-consumption patterns where available. As a result, two experts exposed to the same event cycle can receive different information environments. Source mixes are versioned and audited. The system does not inject a conclusion or a preferred answer. It supplies the context a person in that role would encounter and lets the expert reason from it. That is the working difference between this and a generic prompt to "act as" a customer.

What the literature supports

The approach sits on a substantial and growing research base. Language models conditioned on demographic profiles can act as proxies for human subpopulations (Argyle et al., 2023). They reproduce the aggregate findings of classic human-subject studies (Aher, Arriaga & Kalai, 2023).

A related line of work shows that LLM-backed agents grounded in two-hour interviews with real, named individuals can match those individuals' own survey answers about 85% as well as the individuals match themselves two weeks later (Park et al., 2024).

---

4. How they are tested, and why trust them

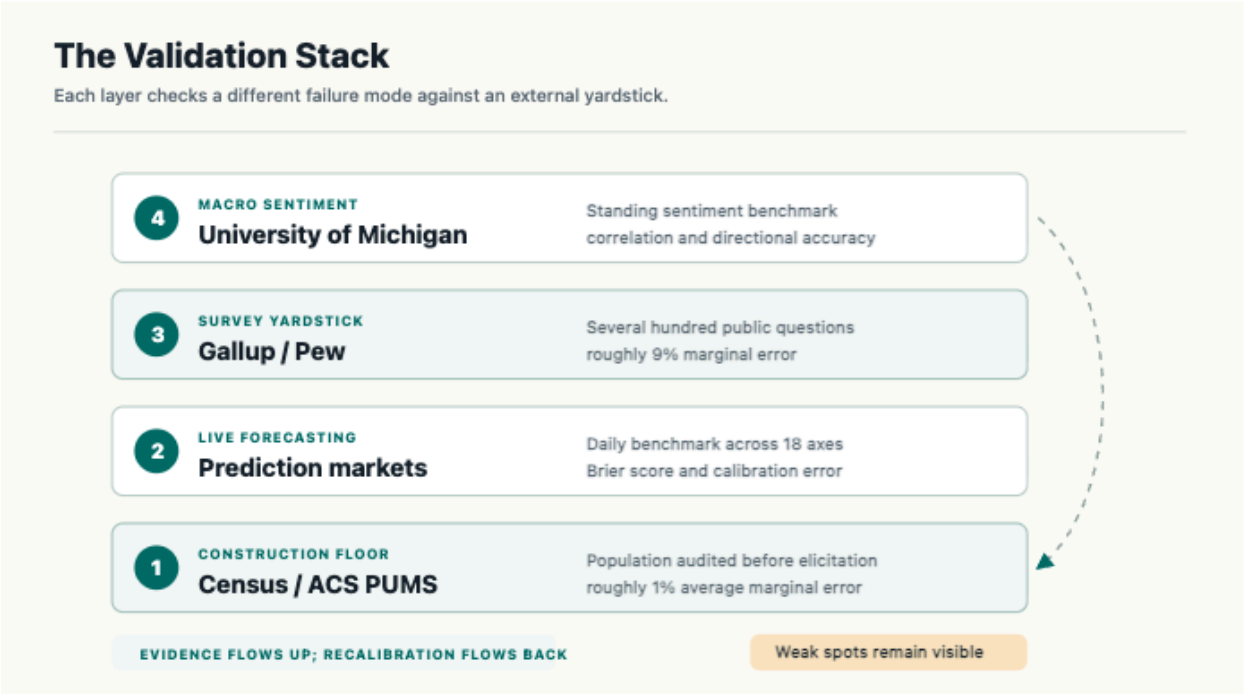
FishDog's validation runs in layers. We measure construction quality, benchmark fit, live forecasting, variance, and drift.

The validation framework is designed around the failure modes a technical or investment-risk reviewer would test: training-data leakage, memory contamination, distributional fidelity, prompt and model drift, variance compression, and benchmark performance over time.

The framework was designed to answer the questions a quant, data-science, or compliance reviewer will ask:

- Does the population match known external marginals before any opinion is elicited?
- Are benchmark questions held out from product tuning where possible?
- Are outputs evaluated on distributions, not only means?
- Are prompt, model, source, and population versions pinned so a re-run is actually a re-run?
- Are training-data leakage and memory contamination structurally controlled?
- Do failure modes produce visible diagnostics rather than disappearing into a single average?

Figure A. The four-layer validation stack, lowest layer first; results feed back into recalibration.



Validation dashboard

Exhibit 4. The validation stack: each layer is benchmarked against an external yardstick, with known weak spots stated alongside the result.

Benchmark layer	Sample / period	Primary metric	Current result	Known weak spots
Census / ACS PUMS construction	Selected demographic and labor-force marginals	Mean absolute marginal error	~1% average marginal error	rare cells, sparse occupations, small geographies
Prediction markets	Daily benchmark across 18 event axes	Brier score, calibration error, resolution accuracy by domain	Stronger on technical and expertise-driven events; weaker on gossip/pop-culture events	low-information events, celebrity and entertainment markets
Gallup / Pew comparison	Several hundred public survey questions	Mean absolute error versus published human results; variance fit where available	~9% marginal error	identity-shaped, highly polarized, or wording-sensitive questions
University of Michigan consumer sentiment	Standing benchmark against published sentiment series	Correlation, directional accuracy, and mean absolute error versus published index	Active monitoring layer	macro turning points, sudden shocks, recession-transition periods

Source: FishDog validation runs, as of June 2026, against U.S. Census and ACS/PUMS, public real-money prediction markets, Gallup and Pew published results, and the University of Michigan Surveys of Consumers. The external instruments are third-party; marginal-error and accuracy figures are FishDog estimates against them.

Layer 1: the construction floor

Before any opinion is elicited, the population is audited against the real one, calibrated to Census and PUMS at roughly 1% marginal error across selected marginals. This layer provides a representativeness check that exists independent of any single question.

Layer 2: prediction markets, daily

Thesis Lab runs its experts against prediction-market events every day, across 18 axes. This provides a signal-versus-outcome test, because a real-money market is the counterparty and the event resolves to a fact. Performance varies by domain: accuracy is strongest on technical and expertise-driven events, for example energy and product questions, and weakest on gossip or low-information events, for example the timing of a celebrity marriage, where marginal error rises.

The 18 axes do more than grade resolved questions. Because benchmark questions are scored on the same axes, the framework builds a map of the question surface area. A new question can be compared with prior questions of similar shape, which gives reviewers an expected reliability range before the output is treated as signal. A near-term geopolitics question carries a different reliability profile than a low-information celebrity question. An analyst can read that profile while there is still time to act, before the event resolves.

For academic support, the paper relies on the established prediction-market literature: the canonical survey of the field (Wolfers & Zitzewitz, 2004) and the long-run accuracy of the Iowa Electronic Markets, a real-money academic market that beat contemporaneous polls 74% of the time across 964 polls (Berg, Nelson & Rietz, 2008). Recent analyses of large commercial markets such as Polymarket are useful contemporary context but are working-paper or journalistic in status and report contested accuracy, as a result we treat them as context rather than as the core evidence.

Layer 3: Gallup and Pew

A few hundred questions drawn from established public surveys are put to the experts and compared against the published human results, at roughly 9% marginal error. This layer benchmarks against a recognized external yardstick rather than a FishDog-defined one. Where published cross-tabs are available, we compare both the aggregate answer and the distribution by demographic group.

Layer 4: University of Michigan consumer sentiment

The University of Michigan Surveys of Consumers have run since 1946 and feed the federal leading-indicator index; they are among the most scrutinized consumer-attitude measures in the world (University of Michigan, Survey Research Center). Thesis Lab benchmarks against the same instrument as a verification layer and a standing connection to a published real-world series.

The feedback loop

These layers feed back into the model. Drift against prediction markets and against Michigan flags where it needs adjustment: weak areas, for example popular culture, get more signal in post-training or source hydration; strong areas, for example politics and economic data, are left alone. Validation is a closed loop for recalibrating against live outcomes. The operating question is whether accuracy holds over time, and the loop is designed to monitor that.

---

5. Known limitations

We are aware of potential failure modes. The main ones are below, paired with the mitigations we use today.

- **Distributional flattening and instability.** The most important critique in the literature: synthetic respondents can land near the right average while showing materially less variance than real humans, can be highly sensitive to prompt wording, and can drift as the underlying model updates (Bisbee et al., 2024). Layered validation and version control are designed to catch this failure mode. We monitor variance against benchmark distributions, not only central tendency, and we control model and prompt versions so that a re-run is a re-run.
- **Demographic misalignment.** Model opinion distributions can be systematically misaligned with specific demographic groups, and steering toward a target group does not fully close the gap (Santurkar et al., 2023). Calibration reduces this; it does not completely eliminate it.
- **Identity-group flattening.** Substituting for human participants is hardest where identity shapes the response, and current training tends to flatten and sometimes mis-portray identity groups (Wang, Morgenstern & Dickerson, 2025). We are most cautious, and most reliant on validation and human triangulation, when positionality drives the answer.
- **Inherited bias.** Language models carry measurable political and social bias from their training data (Feng et al., 2023). For that reason, we benchmark against external ground truth continuously rather than trusting the model's defaults.
- **Training-data leakage.** A model with tool access may have seen the data it is asked to predict, which would turn prediction into reproduction. Our mitigations are structural: daily prediction-market tests run on events that postdate training, benchmark sets are versioned, and the population is calibrated to distributions rather than to memorized individuals.
- **Modeled enrichment error.** The move from census member to expert requires inferred attributes, such as role seniority or procurement exposure, that are not directly observed in PUMS. These enrichments make expert recruitment useful, but they are modeled variables and can be wrong. We disclose sample composition and avoid treating enriched attributes as private facts.

These mitigations reduce the problems, but without removing them. Human market research has its own unfixed failure modes, set out in Section 1. The case for Thesis Lab is that its flaws can be measured and corrected on a loop, and that it avoids several of the specific failures that quietly degrade the human baseline.

---

6. Worked example: enterprise IT budgets, SaaS versus AI versus custom builds

Thesis Lab is deliberately general: the same recruited-population method works at the macro level (consumer stress, tariff and rate sensitivity, regional employment sentiment), at the sector level (GLP-1 impact on food and beverage, AI budget reallocation, EV adoption, travel demand, healthcare staffing), and at the single-name level (recruiting synthetic customers, channel partners, and front-line staff in a company's category to pressure-test demand, switching, and pricing on public signal alone). The worked example below runs one of these in full so the depth and the disclosure discipline are visible end to end. A companion paper collects shorter examples across the other levels; contact us if that would be useful to your team.

The example below illustrates the kind of decision-grade detail a recruited synthetic panel can produce, of a depth an analyst would otherwise assemble over weeks of expert calls. This is an illustrative cut, not a census of the buyer base.

Key takeaway. Enterprise software budgets are roughly flat and reallocating rather than expanding. Artificial-intelligence spend is largely cannibalizing incumbent point-solution software, not adding to it. The clearest disruption short-list is the point-solution layer, while systems of record remain durable.

Method and sample. A Thesis Lab panel of 10 United States enterprise IT decision-makers, including chief information officers, IT and engineering leaders, and systems managers, spanning regulated environments and a conservative public-education outlier. As with any panel, we disclose the sample and avoid extrapolating beyond it.

Exhibit 5. Composition of the illustrative enterprise-IT decision-maker panel (n=10).

Panelist type	Sector	Organization size band	Region	Budget authority
CIO	Healthcare services	1,000-5,000 employees	South	final / shared
VP Engineering	B2B software	250-500 employees	West	final
IT Director	Manufacturing	1,000-5,000 employees	Midwest	recommender / shared
Systems Manager	Public education	500-1,000 employees	Northeast	recommender

Head of Data	Financial services	1,000-5,000 employees	Northeast	shared
Director, Enterprise Apps	Retail / restaurants	5,000+ employees	South	shared
Security and Infrastructure Lead	Healthcare	1,000-5,000 employees	Midwest	recommender
CIO	Professional services	250-1,000 employees	West	final
IT Operations Lead	Logistics	1,000-5,000 employees	South	recommender
Engineering Manager	Industrial technology	250-500 employees	Midwest	recommender

Source: FishDog Thesis Lab synthetic panel, illustrative cut, June 2026. Figures are observed panel outputs, not extrapolated beyond the elicited responses.

What the panel surfaced.

- AI's share of tooling spend grew from low single digits toward the mid-teens, with one decision-maker moving from roughly 2% to roughly 19% year over year.
- Around 60% of the motion was driven by top-down consolidation and return-on-investment mandates rather than bottom-up adoption.
- The build-versus-buy posture shifted from roughly 90/10 toward 80/20, a "build-lite" stance rather than a wholesale move to custom development. One case described a four-week internal build delivering roughly a 60% unit-cost reduction at 90 to 95% of the accuracy of the tool it replaced.
- The most exposed categories, named by the panel, were meeting transcription and notes, writing and grammar tools, lightweight business-intelligence overlays, simple survey and form tools, asynchronous video, and small knowledge-base chat add-ons.
- The most resilient categories were the systems of record: enterprise resource planning, customer relationship management, human-resource information systems, quality and lifecycle management, and student information systems, protected by auditability and the friction of change.
- New procurement gates appeared consistently: pilots on first-party data, "no training on customer data," data-residency and bring-your-own-key requirements, version pinning, logging and role-based access control, spend caps with token metering, and single sign-on.

Dispersion mattered. The signal was not simply "AI budgets up." The more correct, and useful, read was that budgets were flat to modestly up, while the mix changed.

Representative panel language. Short excerpts make the signal more auditable than averages alone:

> "I am not getting a bigger software budget. I am getting permission to cancel three tools if the AI workflow can cover 80% of what they did."
>
> "Systems of record are not where we experiment. We experiment around the edges: notes, reporting overlays, internal search, forms, and support deflection."
>
> "The new gate is not whether the demo is impressive. It is whether legal, security, and finance can see the data path and cap the usage."

Research read. Budgets are flat and reallocating. The point-solution layer in transcription, writing, forms, and BI overlays is the disruption short-list. Systems of record are durable. The likely winners bundle governed AI that replaces two or three point tools at renewal. Every figure above was elicited from the panel and is reported as observed; nothing here is extrapolated beyond what the experts stated.

---

7. How to use Thesis Lab in the analyst workflow

Thesis Lab fits into existing primary-research practice as a leading indicator and a complement, not a substitute.

Figure B. Where Thesis Lab sits in the research workflow, feeding a triangulated read into the investment committee.

- **Before an expert call**, shape and pressure-test the hypothesis, identify which questions are worth a paid human hour, and arrive better briefed.
- **Between prints**, generate a fast read on a sector or cohort when fielding a human survey is too slow or too costly to justify.
- **On any cohort, on demand**, segment into a micro-population that no human panel has been recruited for, and probe it.
- **After a market-moving event**, re-run the same panel the same day to read the change, rather than waiting a quarter for the next wave.

We use it the way we suggest you use it: as a fast, cheap, repeatable signal to triangulate against expert calls and external data, with its provenance and limits stated. It is an input, not an oracle.

Where Thesis Lab does not fit

Thesis Lab is the wrong tool for:

- **Material non-public information, or any private, single-company fact.** It does not possess these by construction, which is the source of its compliance advantage and also a hard limit on what it can tell you.
- **A final investment decision taken in isolation.** It is one input to triangulate, not a verdict.

- **Questions where a specific lived experience or identity is the whole answer.** The literature shows that synthetic respondents can flatten and misportray these cases (Section 5), so we lean hardest on validation and on human research here.
- **Niches with little public signal** to calibrate the population or feed the experts.

---

8. Technical appendix: what a reviewer should be able to inspect

Population frame

- ACS/PUMS vintage used.
- Population size generated.
- Geography level and suppression rules.
- Marginals and joint distributions used for calibration.
- Average, median, and tail marginal error.
- Treatment of rare cells, sparse occupations, and small geographies.

Recruitment and enrichment

- Variables directly observed from the census/PUMS frame.
- Variables inferred or enriched after construction.
- Constraints used for firmographic and role-specific inference.
- Personality model and the number of modelled dimensions beyond the five OCEAN factors.
- Source of personality-trait distributions and how they are audited against reference data.
- Disclosure format for panel composition.
- Rules for rejecting implausible expert profiles.

Source grounding and source control

- Source categories used for general news, regional context, weather, trade publications, filings, and domain literature.
- Update frequency.
- Retrieval versus fine-tuning distinction.
- Source versioning and replayability.
- Rules for excluding client-provided information from global product memory.

Prompt, model, and memory control

- Base model and version.
- Prompt version.
- Population version.
- Source-context version.
- Persistent-memory schema.
- Episodic-memory retention and deletion policy.
- Re-run policy after model or prompt updates.

Benchmarking and leakage controls

- Benchmark question inventory and source.
- Holdout policy.
- Prediction-market event-selection rules.
- Measurement window.
- Metrics: mean absolute error, Brier score, calibration error, variance ratio, directional accuracy, and correlation where appropriate.
- Procedures for detecting training-data leakage.
- Procedures for detecting memory contamination.

Reporting standard

Every client-facing study should disclose:

- Panel size and composition.
- Whether attributes are observed, inferred, or supplied as role-specific source context.
- Date and time of run.
- Model, prompt, population, and source versions.
- Known limitations relevant to the study question.
- Whether results are presented as observed panel outputs or extrapolated estimates.

---

9. Conclusion

The inputs the market relies on for primary research are getting noisier and more expensive. Expert networks converge and get gamed; survey panels self-select and get coached. A synthetic method that wants to be taken seriously should not claim to equal real people. It should build from a calibrated population, recruit experts rather than invent them, benchmark against real-world ground truth, and state its limits. That is the standard we hold Thesis Lab to.

References

Aher, G., Arriaga, R. I., & Kalai, A. T. (2023). Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. ICML 2023, PMLR 202, 337-371.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of One, Many: Using Language Models to Simulate Human Samples. Political Analysis, 31(3), 337-351.

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.

Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., & Larson, J. M. (2024). Synthetic Replacements for Human Survey Data? The Perils of Large Language Models. Political Analysis, 32(4), 401-416.

Feng, S., Park, C. Y., Liu, Y., & Tsvetkov, Y. (2023). From Pretraining Data to Language Models to Downstream Tasks. ACL 2023, 11737-11762.

Goldberg, L. R. (1990). An Alternative "Description of Personality": The Big-Five Factor Structure. Journal of Personality and Social Psychology, 59(6), 1216-1229.

John, O. P., & Srivastava, S. (1999). The Big Five Trait Taxonomy. In Handbook of Personality (2nd ed., pp. 102-138). Guilford Press.

Little, R. J. A. (1993). Post-Stratification: A Modeler's Perspective. Journal of the American Statistical Association, 88(423), 1001-1012.

McCrae, R. R., & Costa, P. T. (1987). Validation of the Five-Factor Model of Personality Across Instruments and Observers. Journal of Personality and Social Psychology, 52(1), 81-90.

Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., Willer, R., Liang, P., & Bernstein, M. S. (2024). Generative Agent Simulations of 1,000 People. arXiv:2411.10109v1.

Pew Research Center (2018). How Different Weighting Methods Work.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose Opinions Do Language Models Reflect? ICML 2023, PMLR 202, 29971-30004.

U.S. Census Bureau. American Community Survey and Public Use Microdata Sample (PUMS).

University of Michigan, Survey Research Center, Institute for Social Research. Surveys of Consumers (Index of Consumer Sentiment), founded 1946.

Wang, A., Morgenstern, J., & Dickerson, J. P. (2025). Large Language Models That Replace Human Participants Can Harmfully Misportray and Flatten Identity Groups. Nature Machine Intelligence.

Wolfers, J., & Zitzewitz, E. (2004). Prediction Markets. Journal of Economic Perspectives, 18(2), 107-126.

Disclosures

This document is provided for institutional and professional investors for informational purposes only. It is not investment advice and is not a recommendation to buy or sell any security. FishDog is not a registered investment adviser. Synthetic-persona research is a complement to, not a replacement for, other primary research and should be triangulated against independent sources. Thesis Lab is a FishDog product launched in partnership with Eagle Alpha; FishDog is responsible for the methodology and claims in this paper.

FishDog methodology standard: this document commits to the validation framework described in Section 4 and to disclosing sample composition and known limitations on every study. Figures attributed to panels are reported as observed and are not extrapolated beyond the elicited responses unless explicitly labeled as modeled estimates.
