

AI POPULATION SIMULATION

FORWARD-GRADED POPULATION SIMULATION WITH REACTIVE COUNTERFACTUALS

The category validates itself by reproducing answers it already has. This paper reports what happens when you grade the same machinery **forward** — and publishes the results that fail.

AGENT-N RESEARCH · DRAFT 2026-08-15 · WHAT IS PROVEN, WHAT IS NOT, AND HOW TO CHECK



ABSTRACT

Agent-N forecasts how real electorates behave and how they react to interventions. Unlike the prevailing AI-simulation approach — which validates by **retrodiction**, reproducing a survey whose answers already exist — Agent-N is graded **forward and out-of-sample**, with no outcome leakage. On the CA 2026 Governor primary it achieves held-out Spearman **0.961 forward** (poll-seeded, no outcome touches it) against **0.961 in retrodiction mode** — a near-zero drop that distinguishes a transferable structure from a memorized fit — and the learned structure generalizes cold to three unseen statewide races at **2.51pp out-of-race**.

Beyond forecasting, an evolving agent panel produces a **reactive coupling** effect — the trailing side counter-mobilizes under an issue shift — that survives an adversarial statistical battery (permutation $p = 2 \times 10^{-4}$, 99% CIs excluding zero) and that a static model structurally cannot produce. We are equally explicit about the negative results: the stronger per-voter individuation claim **fails** its cleanest test, and the point-forecast layer **cannot** beat a liquid prediction market. We separate these on two scoreboards and claim only what each earns.

01 THE PROBLEM

Surveys and focus groups are slow, expensive, and backward-looking; people are poor predictors of their own behavior, and stated preference is not revealed preference. AI population simulation promises a faster substitute, and the category leader (Aaru, roughly a \$1B headline valuation) has real enterprise traction.

But the category's public credibility rests almost entirely on **retrodiction** — reproducing known survey results and reporting a correlation — with no published held-out forward metric, no whitepaper, and, in the one independently gradeable forward case, the 2024 US presidential, a **miss**.



AN EXPENSIVE ROOM AND A SLOW ANSWER — THE METHOD THE PAPER ARGUES PAST

THE OPEN GROUND

Honest forward, graded, pre-registered forecasting is unclaimed. That is what this paper documents — including the parts of it we could not make work.

02 TWO SCOREBOARDS

We hold two claims strictly separate and never let one launder credibility onto the other. Every result below is filed under exactly one scoreboard, with its own evidentiary bar.

SCOREBOARD 1 — ACCURACY

Point forecasts of who wins and what share, graded against certified results, out-of-sample.

TABLE STAKES

SCOREBOARD 2 — COUNTERFACTUAL

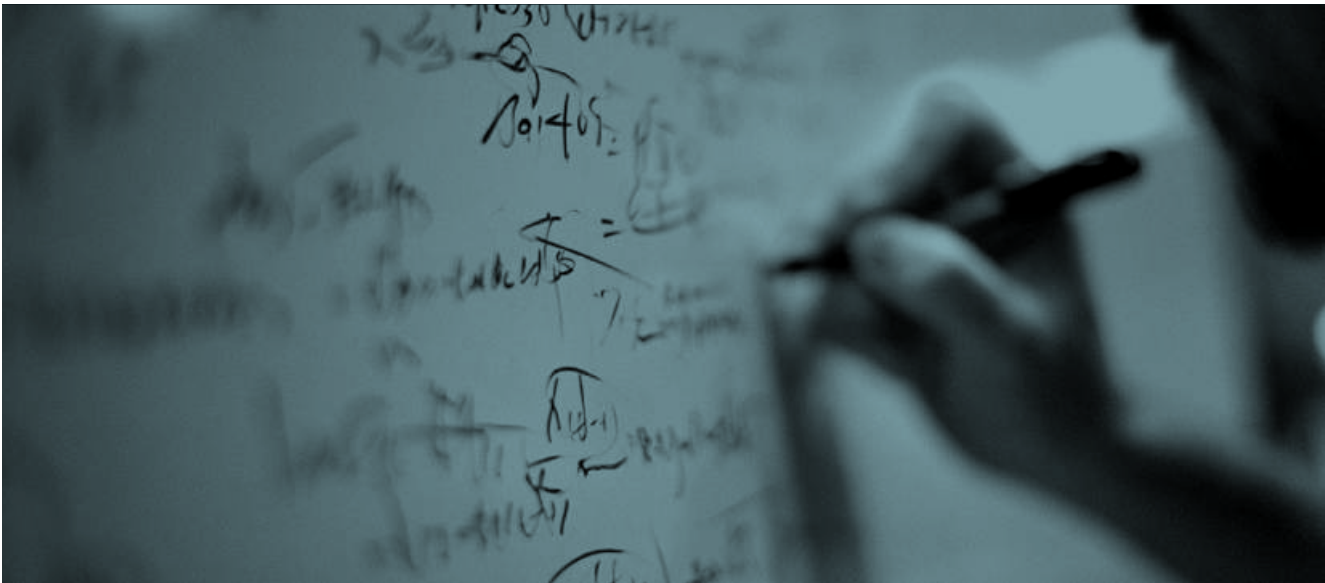
What moves if an input changes — reactive dynamics a static model cannot produce.

THE DIFFERENTIATOR

FIGURE 1 — THE TWO SCOREBOARDS, HELD STRICTLY APART

03 FORWARD-GRADED ACCURACY

METHOD



A faithful electorate is built from PUMS joint microdata, county-faithful subsampling, then raked (IPF) to certified registration and CPS turnout; margins land at roughly 0.00pp. The vote is **not** taken by an LLM — argmax over an LLM's priors is uncalibrated and politically biased, and it put Steyer at 4% in Marin, where he won at 35.9%. Instead a **calibrated conditional-logit layer** votes over the faithful population using sourced features. Candidate appeal is seeded two ways:

- **Retrodiction mode** — appeal seeded from the known statewide result, structure fit on train counties, reproducing the held-out county breakdown. This is the industry-standard test.
- **Forward mode** — appeal seeded from a pre-election poll only. No county outcome touches appeal at any point.

RESULTS — CA GOVERNOR PRIMARY 2026

8 candidates, 17 held-out counties = 136 county×candidate cells. `scripts/demos/retrodiction-vs-forward-demo.ts`

MODE	SPEARMAN (PRED VS ACTUAL)	MAE
Retrodiction	0.961	2.50pp
Forward (poll-seeded, no leakage)	0.961	2.88pp

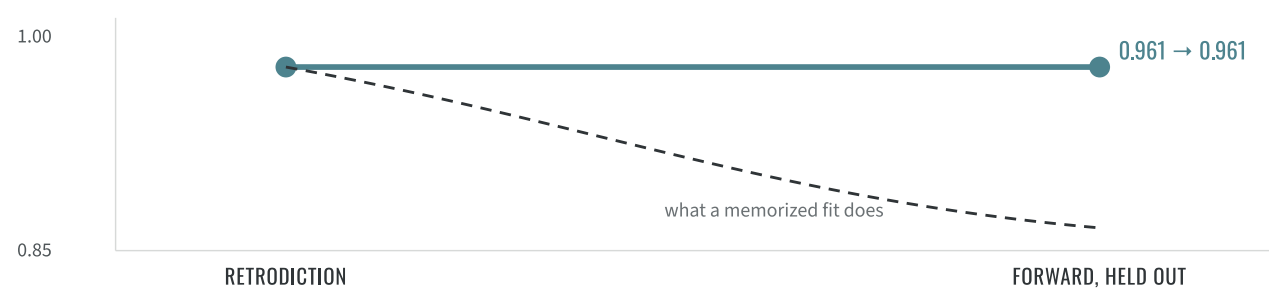


FIGURE 2 — THE NEAR-ZERO DROP IS THE RESULT

The **near-zero drop** is the key finding. A model that memorizes the outcome aces retrodiction and collapses forward. Agent-N does not, because the county-bending signal is a real, transferable relationship rather than a fit to the answer.

GENERALIZATION

Freezing the generic partisan structure learned on Governor and applying it cold to three unseen 2026 statewide races — Insurance Commissioner, Treasurer, Superintendent — yields **2.51pp out-of-race**, beating the naive statewide-everywhere baseline on all three and hurting none (`scripts/portfolio-cross-race-test.ts`). Best-case held-out, with outcome-learned appeal, is 1.52pp.

0.961

Held-out forward Spearman, poll-seeded

2.88_{pp}

Forward MAE on the same run

2.51_{pp}

Out-of-race, three unseen statewide contests

AGAINST THE FAIR BASELINE

Statewide-everywhere is a strawman. The fair baseline is **uniform swing** — the prior governor general's county margins plus the statewide swing. It is a strong baseline, because county partisan geography is highly stable, and we do **not** dominate it. On a two-cycle head-to-head (`scripts/demos/governor-baseline-comparison.ts`) Agent-N beats it in 2018 (Spearman 0.983 vs 0.946; MAE 5.47 vs 9.67pp) and loses narrowly in 2022 (0.981 vs 0.991; 4.33 vs 3.26pp): a 1–1 split.

Two honest corollaries follow. Registration-only performs about as well as the full structure — the demographic terms add little, and this is roughly 90% partisan geography. And uniform swing wins when the swing is uniform, while our method wins when the environment shifts.

THE DEFENSIBLE EDGE

The claim is not that we beat the baseline. It is that our method **needs no comparable prior race**. Uniform swing is impossible forward — there is no prior Becerra-vs-Hilton — and impossible on new or down-ballot races, where no prior same contest exists. That is exactly where prior-free structural transfer is the only option that works, and where it delivered the 2.51pp above. On races with a clean prior we are merely competitive; the value is in the majority of races, and all forward forecasts, that have none.

DISCIPLINE



WRITE THE BET DOWN BEFORE YOU SEE THE RESULT

Held-out counties never enter the fit. Forward mode is poll-seeded, never outcome-seeded. Features are sourced, never LLM-authored. Actuals are scoring-only and never enter agent state — a thermometer, not a thermostat.

A NEGATIVE WE PUBLISH

The forward statewide topline approximates the seeding poll, and where the poll misreads a late surge, so does Agent-N. Aggregating the county model back to a statewide number is roughly **2× worse than the raw poll** (2.59 vs 1.33pp mean absolute error) and gets the pivotal top-two call wrong, because the county structure imposes a partisan-geography prior that fights an out-party surge.

IMPLICATION

Agent-N **cannot** beat a liquid prediction market on the binary "who wins the state" question. The county edge is cross-sectional — where the vote is, not how much of it there is in total. That edge is a campaign product, not a trading one. `scripts/market-backtest/statewide-vs-poll-check.ts`



THE CROSS-SECTIONAL QUESTION: WHERE THE VOTE IS

04 REACTIVE COUNTERFACTUALS



THE SECOND STAGE: THE OTHER SIDE MOVES TOO

MECHANISM

Agents never vote. An intervention moves the inputs the calibrated layer reads, and the same Layer-1 model re-votes. The evolving panel adds a second stage: under a hypothetical issue-salience rise — framed as an explicit hypothetical, so no false fact enters forward-only state — the trailing camp **counter-mobilizes**. The moat statistic is reactive — placebo per agent. Both are second-stage draws; they differ only by the opponent's counter-move. A static coefficient bump has no second stage, so it produces this gap = 0 by construction.

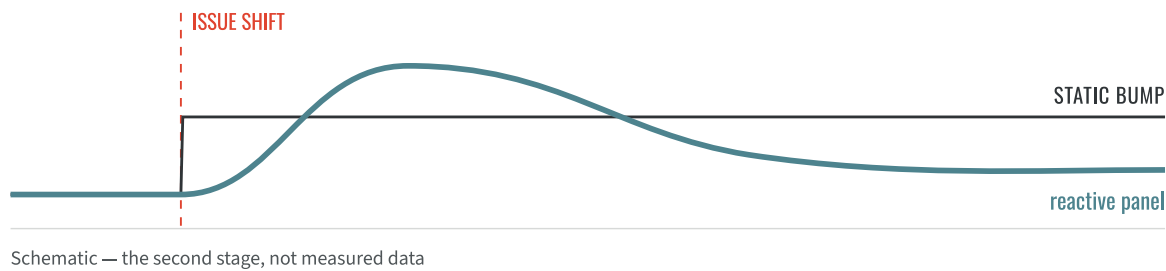


FIGURE 3 — COUNTER-MOBILIZATION AND DECAY, WHICH A COEFFICIENT BUMP CANNOT PRODUCE

ADVERSARIAL VALIDATION

The instrument is self-validated: it detects a coupled panel at $p = 2 \times 10^{-4}$ and refuses a null panel at $p = 0.87$ (scripts/counterfactual/coupling-stress-test.ts). On both $N=120$ runs, the main effects survive a full skeptic's battery — paired sign-flip permutation, bootstrap CIs at 90/95/99%, jackknife leave-one-out, 10% trimming, and a sign test.

EFFECT	PERMUTATION P	99% CI	JACKKNIFE	AGENTS ON SIDE
Steyer +1.08	2×10^{-4}	excludes 0	sign-stable	78%
Hilton -1.04	2×10^{-4}	excludes 0	sign-stable	79%
Becerra +0.76 (wildfire)	2×10^{-4}	excludes 0	sign-stable	76%

No single agent drives the sign, and the effect survives trimming the tails. One small effect, primary-Becerra -0.22, is marginal — $p = 0.037$, with the 99% CI including zero — and we do not lean on it. The statistic is κ -invariant, a sign and gap rather than a magnitude, so it stands without fitting the leaning-to-vote scale.

THE NEGATIVE WE PUBLISH

The stronger claim — that each agent is an individuated mind whose reasoning separates by its seed — **fails**. A pre-registered, self-validated within-vote position-recovery test, whose instrument fires at $z = 19$ on synthetic signal, reads the real panel at **$z = 1.24$: theater** (scripts/demos/same-party-within-vote-recovery.ts). This is the second failure of per-voter individuation, after the persona-strip null.

WHAT WE THEREFORE DO NOT CLAIM

We do not claim conscious individuated agents. The demonstrated, defensible value of the substrate is **aggregate reactive coupling**, full stop.

WHAT IS NOT YET PROVEN

The coupling is an **in-sample mechanism**: the panel genuinely does something a bump cannot. It has not been sealed out-of-sample against a resolved forward outcome, because the clean pre-shock instrument that would isolate it — county-level contestation — is not publicly available. It is naturally a with-customer, design-partner

artifact. The out-of-sample counterfactual seal is the central open item of this research programme.

05 RELATION TO PRIOR WORK

- **Aaru and silicon sampling.** Retrodiction plus self-reported election PR, with no held-out forward metric and no methodology paper. Agent-N occupies the gap: graded forward, reproducible, with a counterfactual mechanism the incumbent explicitly disclaims.
- **Academic caution on LLMs-as-humans** (PNAS 2025; the distributional-collapse results) is well founded — and is why Agent-N removes the LLM from the calibrated vote, confining the evolving substrate to the reactive-coupling role where it earns its keep, rather than asking an LLM for a calibrated quantity.



AFTER THE COUNT

06 HONEST SCORECARD

CLAIM	STATUS
Forward-graded accuracy (county level)	Proven — out-of-sample and out-of-race
Beating a liquid market on statewide "who wins"	Disproven — topline $\approx$ poll, roughly 2 $\times$ worse
Reactive coupling (aggregate)	Proven in-sample, adversarially validated
Per-voter individuation	Disproven — theater
Out-of-sample counterfactual seal	Open — needs a resolved bet and design-partner data

Two of five rows are negative. That is the point of publishing the scorecard: a claim is only worth what its hardest test says, and the tests that failed are named as plainly as the ones that passed.

07 REPRODUCE EVERYTHING

```
npx tsx scripts/demos/retrodiction-vs-forward-demo.ts # 0.961 retro / 0.961 forward
npx tsx scripts/portfolio-cross-race-test.ts # 2.51pp out-of-race
npx tsx scripts/counterfactual/coupling-stress-test.ts # coupling p=2e-4, survives battery
npx tsx scripts/demos/same-party-within-vote-recovery.ts # individuation THEATER
npx tsx scripts/market-backtest/statewide-vs-poll-check.ts # topline ~2x worse than the poll
```

Companion artifacts: EVOLUTION_LOG.md (primary record), docs/DEMO-retrodiction-vs-forward-vs-aaru.md, docs/RESEARCH-aaru-competitive-intel-2026-08-15.md, and the same-party demo record (internal).



THE ROOM THIS PAPER IS WRITTEN FOR

Run the numbers yourself. Every figure in this paper reproduces from the repository in one command — **which is the whole argument.**