MM-TR-001 TECHNICAL REPORT JULY 2026 SINGAPORE

Calibrating Synthetic Society

Census-grounded agent populations that disagree like real ones.

Motif Motion Research · Motif Motion Pte. Ltd., Singapore · hello@motifmotion.sg

WORKING PAPER · V0.1 · DRAFT FOR COMMENT

ABSTRACT

Language models can role-play survey respondents well enough that "silicon sampling" is now a serious research instrument — and badly enough that an uncalibrated synthetic panel is a confident fabrication machine. This report describes how CrowdOS, our population-intelligence platform, approaches the problem as one of *calibration* rather than imitation. We describe the system's design — a persistent panel of 4,000+ OCEAN-modelled archetypes spanning 21 markets, weighted to census margins and run on multi-provider model ensembles — and the evaluation discipline behind it: parity measurement against national survey instruments (Pew, ANES, GSS, OpinionQA), drift monitoring, and explicit instrumentation against sycophancy, the tendency of model-based agents to collapse into agreeable consensus. Our current headline figure is 91.6% top-line parity against matched Pew reference items. We state what that number does and does not mean, the conditions under which synthetic panels break, and why we treat preserved disagreement — not smoothed consensus — as the core signal of a well-calibrated synthetic society.

Keywords: synthetic populations · silicon sampling · survey calibration · sycophancy · LLM agents · census grounding

1 Introduction

A growing body of work shows that large language models, conditioned on demographic personas, can reproduce human survey response patterns with surprising fidelity [1, 3, 4].

The promise is obvious: opinion measurement and stimulus testing at a fraction of the cost and latency of fieldwork. The peril is equally obvious: a language model will answer *any* question in *any* persona with equal confidence, whether or not the underlying distribution it implies bears any relationship to a real population [2, 9].

Our position is that the difference between the two outcomes is not model scale but **calibration discipline**. A synthetic panel is a scientific instrument. Like any instrument, it is only as good as its reference standards, its known error modes, and the honesty of its operators about what it cannot measure. CrowdOS is our attempt to build that instrument properly — and this report is the first in a series documenting how, what we measure, and where it breaks.

2 Related Work

Argyle et al. demonstrated that GPT-3 conditioned on demographic backstories exhibits "algorithmic fidelity" to subgroup response distributions in American electoral surveys [1]. Aher et al. replicated classic human-subject studies with simulated participants [3], and Horton framed language models as "homo silicus" for economic experiments [4]. Park et al. showed that generative agents can sustain believable long-horizon social behaviour [5] and, more recently, that interview-grounded agents can simulate specific individuals' survey responses at meaningful accuracy [6].

The cautionary literature is just as important to us. Santurkar et al. showed that base models reflect particular demographic opinions far more than others [2]; Dominguez-Olmedo et al. showed that survey responses from language models carry systematic biases that naive prompting does not remove [9]. Perez et al. and Sharma et al. documented sycophancy — models telling users what they appear to want to hear — as a persistent, training-induced behaviour [7, 8]. Grossmann et al. situate all of this within a broader transformation of social-science method [10]. CrowdOS is built on the working assumption that *both* literatures are right: the fidelity is real, and so are the failure modes.

3 The CrowdOS Population Model

Figure 1 sketches the platform end to end: a census-grounded archetype panel, administered through a multi-provider model ensemble, with every response cycle feeding a parity-and-drift loop that recalibrates the panel against reference instruments.

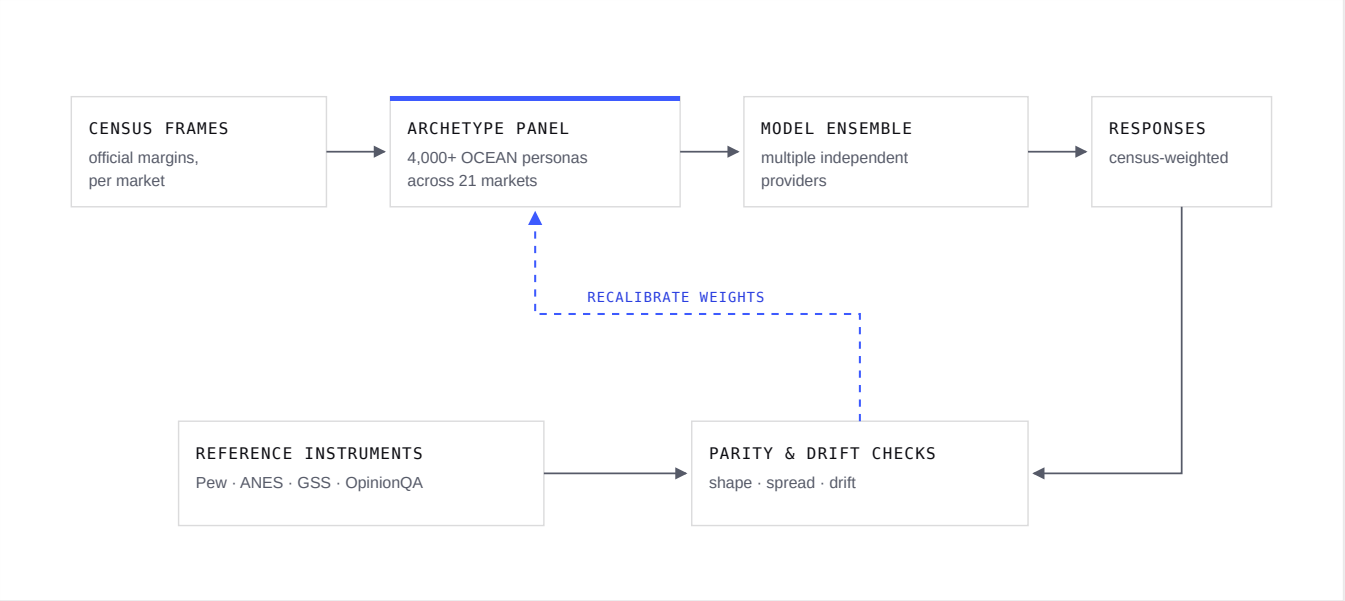


Figure 1 — The CrowdOS calibration loop. Census frames define the archetype panel; a multi-provider ensemble administers studies; responses are continuously checked against reference instruments, and deviations recalibrate the panel's weights.

3.1 Persistent archetypes

CrowdOS does not spin up disposable personas per query. The platform maintains a panel of **4,000+ persistent archetypes** across **21 markets**, each specified by demographic attributes aligned to census categories and a stable personality profile expressed in the five-factor (OCEAN) framework [11]. Persistence matters for two reasons. First, it makes panels *reproducible*: the same question asked twice meets the same society. Second, it makes drift *measurable*: a persistent archetype that answers differently across model versions is a signal we can detect, attribute, and correct.

3.2 Census grounding

Archetypes are sampled and weighted so that panel margins track official census distributions in each market — the same post-stratification logic survey researchers apply to human panels. The platform's population frame is designed to scale to the roughly eight billion people the

world census data describes; any given study draws a weighted panel from that frame appropriate to its target population.

3.3 Multi-provider ensembles

Every model family carries its own opinion profile [2]. A panel simulated by a single model therefore inherits a single provider's bias as a hidden confound. CrowdOS runs archetypes across **ensembles of models from multiple providers**, so that no single model's dispositions masquerade as the population's. Disagreement between providers on the same archetype is itself diagnostic: it flags items where the synthetic signal is model-driven rather than persona-driven, and those items are treated with correspondingly lower confidence.

3.4 Research surfaces

Three product surfaces sit on the population model — Pulse (opinion measurement), Screening Room (stimulus testing), and Ads Labo (advertising performance) — plus a public API and MCP interface. From a research standpoint they are the same instrument pointed at different questions, and they share one calibration substrate.

4 Calibration Methodology

4.1 Reference instruments

We validate against the surveys that population science itself treats as reference standards: **Pew Research Center** studies, the **American National Election Studies (ANES)**, the **General Social Survey (GSS)**, and the **OpinionQA** benchmark derived from Pew's American Trends Panel [2]. These provide professionally fielded, publicly documented human distributions with known sampling frames.

4.2 Parity

By **parity** we mean agreement between the synthetic panel's response distribution and the human reference distribution on matched items, administered to the synthetic panel under matched question wording and response options, with panel weights matched to the survey's target population. Our headline metric is top-line parity — agreement on the aggregate

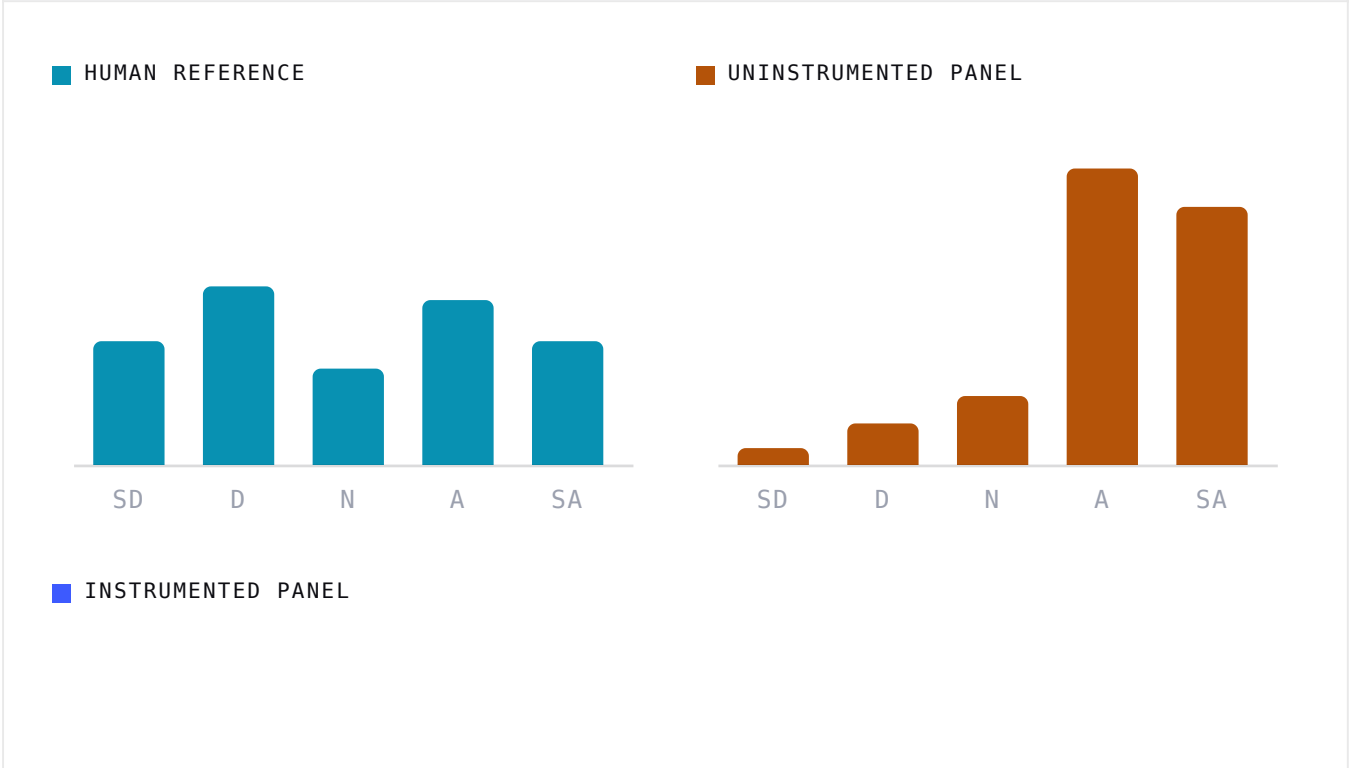
distribution — with subgroup-level parity tracked internally as the harder and more informative target. We report the current headline figure in §6; per-instrument and per-market breakdowns, with the full metric definition, are being prepared for release alongside this series.

4.3 Drift monitoring

Calibration is not a one-time certification. Underlying models are updated by their providers; real populations move; question contexts shift. Persistent archetypes are re-administered reference items on a rolling basis, and deviations beyond tolerance trigger re-weighting or re-grounding of the affected panel segments. A synthetic panel that was calibrated last quarter is, by default, assumed stale until re-measured.

5 Anti-Sycophancy Instrumentation

The failure mode we consider most dangerous in synthetic research is not random error but **flattery at scale**. Language models are trained in ways that reward agreeable answers [7, 8]. Put a concept in front of an uninstrumented synthetic panel and it will, on average, like it — and like it more if the question's framing hints that the asker hopes it will. A panel that smooths away disagreement is worse than useless for research: it manufactures false confidence with the aesthetics of evidence.



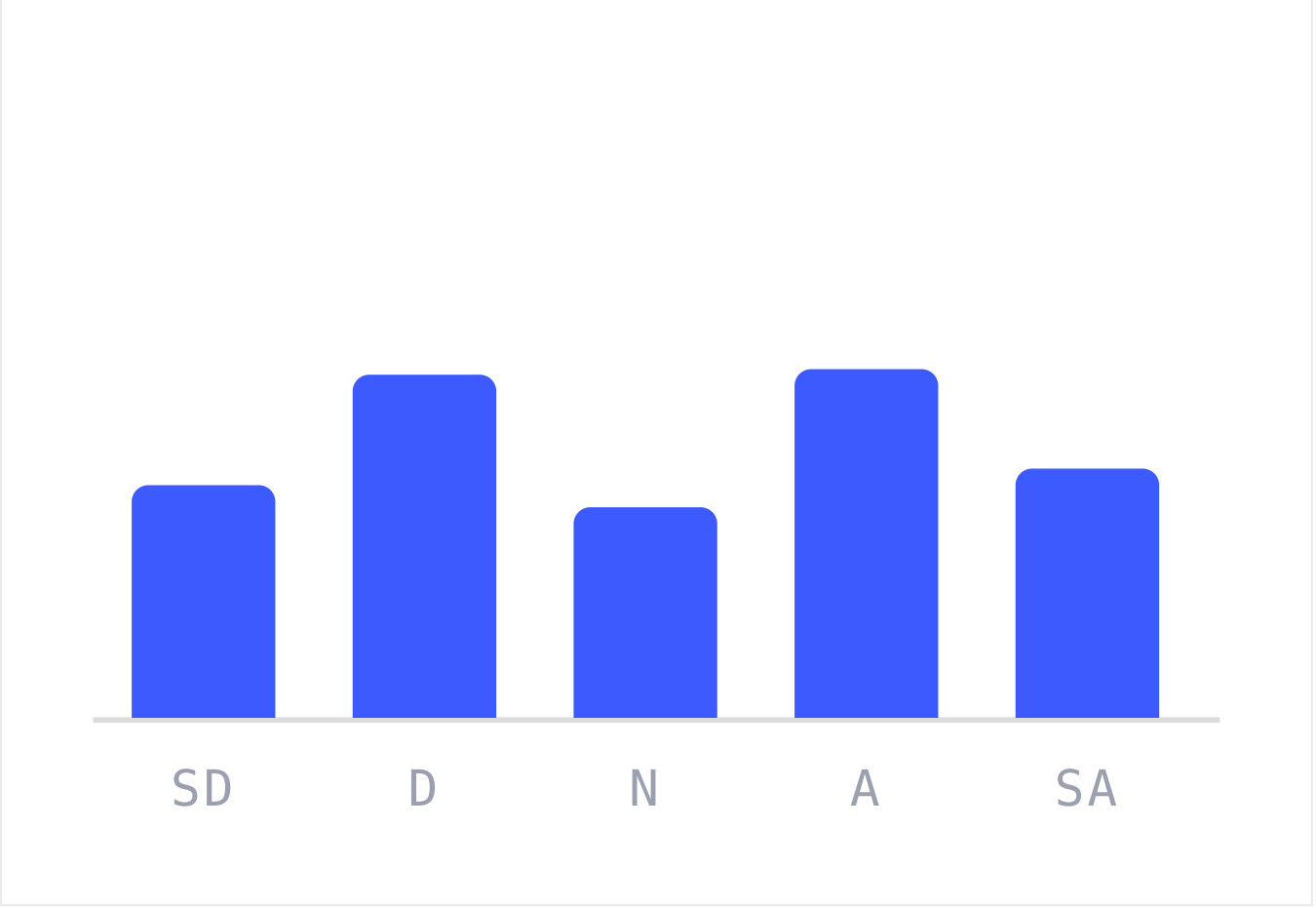


Figure 2 — The sycophancy failure mode. An uninstrumented panel collapses toward agreeable consensus; an instrumented panel preserves the reference spread, including its dissenting minorities. Distributions are schematic — they illustrate the failure mode's shape, not measured data.

CrowdOS treats **disagreement as a calibration target**, not noise. Concretely, the platform:

- **Measures spread, not just central tendency.** Parity checks compare the shape of response distributions — including the size of dissenting minorities — against reference data, so a panel that converges on agreeable consensus fails calibration even when its mean looks right.
- **Shields panels from asker intent.** Study questions are administered to archetypes in neutralised form, decoupled from the client's framing, so that an agent cannot infer — and therefore cannot flatter — what the researcher hopes to hear.
- **Preserves persona friction.** Archetypes carry stable dispositions that legitimately conflict with many stimuli. Instrumentation monitors the rate at which archetypes produce negative, critical, or refusing responses, and flags panels whose criticism rate collapses below reference-consistent levels.

- **Uses ensemble disagreement as an alarm.** When multiple providers' models, given the same archetype, converge unusually hard on praise, the item is flagged for sycophancy review rather than reported at face value.

6 Headline Result, and What We Do Not Claim

Under the methodology of §4, CrowdOS's current headline figure is **91.6% top-line parity against matched Pew reference items**. We publish that number on our product surfaces, and this report exists in part to give it an honest frame.

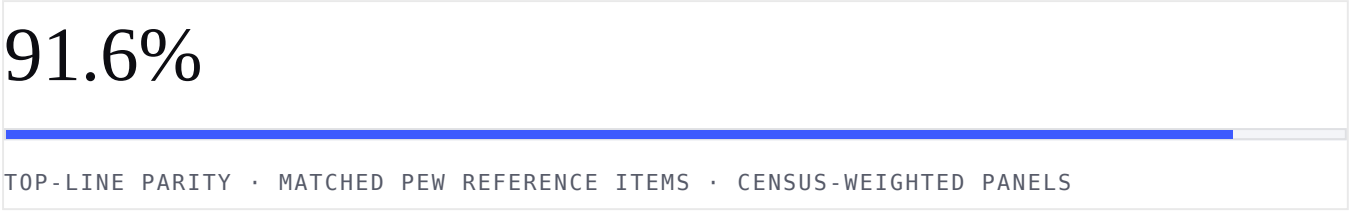


Figure 3 — Current headline parity. Aggregate response-distribution agreement on reference items under matched wording; per-instrument and per-market breakdowns forthcoming (§6).

What it means: on the reference items we administer, the synthetic panel's aggregate response distributions agree with Pew's human distributions at that level, under matched wording and census-weighted panels. What it does *not* mean:

- It is not a claim about **every topic**. Parity is measured on reference instruments; novel stimuli — the questions clients actually bring — inherit calibration indirectly, and we say so.
- It is not a claim about **every subgroup**. Aggregate parity can mask subgroup error, and the literature is clear that models represent some populations better than others [2, 9]. Subgroup parity is the harder target we track internally.
- It is not a claim of **equivalence to fieldwork**. A synthetic population is a model of a population, not the population. For decisions that put people at risk, synthetic panels are a complement to human research, not a substitute.

> Forthcoming in this series: the full parity metric specification, per-instrument and per-market breakdowns, and our sycophancy stress-test protocol with results across model providers.

7 Limitations and Responsible Use

Three limitations bound everything above. **Distribution shift:** reference surveys describe the recent past; populations move, and synthetic panels calibrated to them lag reality in ways drift monitoring can detect but not eliminate. **Representation:** archetypes are built from the categories census and survey data measure; people and positions poorly covered by those instruments are poorly covered by ours, and aggregate parity does not repair that. **Provenance:** model providers' training data and update schedules are outside our control; ensembles hedge this exposure but do not remove it.

On use: CrowdOS is built for research — measuring opinion, testing stimuli, exploring message resonance. It is not built for, and we do not condone, fabricating grassroots sentiment, impersonating real individuals, or presenting synthetic responses as human ones. Synthetic research should be labelled as synthetic wherever its outputs travel.

8 Future Work

Two directions dominate our roadmap. First, **domain-specific weights**: moving population competence from the harness into trained models of our own, beginning with the population-simulation corpus (our Track 01). Second, **open evaluation**: releasing enough of the parity methodology that our headline numbers can be checked rather than trusted. A lab whose premise is that unmeasured systems shouldn't ship owes the field the instruments to measure it by.

References

[1] 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).

[2] Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose Opinions Do Language Models Reflect? *Proceedings of ICML*.

[3] Aher, G., Arriaga, R. I., & Kalai, A. T. (2023). Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. *Proceedings of ICML*.

[4] Horton, J. J. (2023). Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? *NBER Working Paper*.

[5] Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative Agents: Interactive Simulacra of Human Behavior. *Proceedings of UIST*.

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

[7] Perez, E., et al. (2022). Discovering Language Model Behaviors with Model-Written Evaluations. *arXiv preprint*; Findings of ACL 2023.

[8] Sharma, M., et al. (2023). Towards Understanding Sycophancy in Language Models. *ICLR 2024*.

[9] Dominguez-Olmedo, R., Hardt, M., & Mendler-Dünner, C. (2023). Questioning the Survey Responses of Large Language Models. *arXiv preprint*; NeurIPS 2024.

[10] Grossmann, I., Feinberg, M., Parker, D. C., Christakis, N. A., Tetlock, P. E., & Cunningham, W. A. (2023). AI and the Transformation of Social Science Research. *Science*, 380(6650).

[11] McCrae, R. R., & John, O. P. (1992). An Introduction to the Five-Factor Model and Its Applications. *Journal of Personality*, 60(2).

CITE THIS REPORT

```
@techreport{motifmotion2026calibrating,
  title       = {Calibrating Synthetic Society: Census-Grounded Agent
                 Populations that Disagree Like Real Ones},
  author      = {{Motif Motion Research}},
  institution = {Motif Motion Pte. Ltd.},
  number      = {MM-TR-001},
  year        = {2026},
  month       = {7},
  address     = {Singapore},
  url         = {https://motifmotion.sg/papers/calibrating-synthetic-society.html}
```

```
  note         = {Working paper, v0.1}
}
```