MM-TR-003 TECHNICAL REPORT AUGUST 2026 SINGAPORE

Distribution Learning Is Not Respondent Simulation

A pre-registered negative result: fine-tuning on survey microdata failed to transfer to simulated survey respondents.

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

WORKING PAPER · V0.2 · ADDENDUM: AUG 28, 2026

ABSTRACT

We test whether fine-tuning on human survey microdata — a method with demonstrated success at predicting subpopulation opinion distributions (SubPOP; Suh et al., 2025) — transfers to the deployed task of simulating individual survey respondents. We fine-tuned a ~9-billion-parameter open-weight instruct model on 142,052 examples curated from Pew American Trends Panel raw microdata and evaluated it, under a pre-registered criterion, against a heterogeneous five-model frontier ensemble on a 40-item bank of real polling questions — both arms raw, under an identical elicitation harness. The fine-tuned model reached near-floor eval perplexity in its own training format (**1.75**), indicating the target distributions were genuinely learned — yet scored **46.92pp mean absolute error** against the ensemble's **12.53pp**, with **8.6% directional parity**, a **−0.595** cross-item correlation, and **16.3%** of its simulated respondents producing unparseable output. The failure is one of *format transfer*: distribution prediction over demographic subgroups did not transport to individual-level, in-character, structured response generation, and the narrow completion objective displaced the base model's instruction-following within the domain. We analyze the failure, discuss why heterogeneous prompted ensembles are a hard baseline for task-specific fine-tunes in this category, and argue that proprietary-model accuracy claims should be evidenced by pre-registered comparisons of exactly this shape.

Keywords: fine-tuning · negative result · pre-registration · silicon sampling · format transfer · synthetic respondents

1 Background

CrowdOS runs simulated-respondent research: panels of 20–1,000+ AI respondents answer questions independently or debate to consensus. The production system seats each panel across an **ensemble of five frontier reasoning models from different providers**, behind a persona and elicitation layer that is identical for every model. End-to-end accuracy is reported on our public benchmark ledger.

A recurring question in this category — raised by the literature's own successes — is whether a task-specific fine-tuned model should replace prompted frontier models. SubPOP (Suh et al., 2025) demonstrated large gains in subpopulation opinion-distribution prediction from fine-tuning on survey microdata [5]. If that skill transferred to respondent simulation, a small tuned model could plausibly rival much larger prompted ones on this domain. We designed a pre-registered experiment to test exactly that hypothesis, with the decision criterion fixed before training began.

2 Related Work

Silicon sampling. Argyle et al. (2023) showed that a language model conditioned on demographic backstories can reproduce aggregate response distributions of human subpopulations with surprising fidelity — the result that seeded this product category [1]. Santurkar et al. (2023) formalized measurement of opinion alignment against Pew's American Trends Panel and introduced the steering format much of the later literature builds on [2].

Individual-level simulation. Park et al. (2023) demonstrated believable long-horizon generative agents [3]; Park et al. (2024) grounded 1,000 individual agents in two-hour qualitative interviews and reproduced held-out survey responses at the individual level [4]. The important contrast for our purposes: these systems simulate *people*, not *marginals*.

Fine-tuning on survey distributions. SubPOP fine-tunes language models on large banks of (question, subpopulation) → response-distribution pairs curated from ATP microdata, substantially improving subpopulation distribution prediction, including transfer to unseen surveys [5]. Its released pipeline (BSD-3) is the recipe we replicated.

Our question sits between these literatures: **does the distribution-prediction skill that SubPOP-style training demonstrably produces transport into a production respondent simulator — an individual-level, in-character, structured-output task?**

3 Method

3.1 Training data

We trained on **Pew Research Center American Trends Panel (ATP) raw microdata**, obtained under Pew's download agreement (a personal, revocable grant; the files never enter a repository and are never redistributed). The packaged third-party release of the SubPOP dataset is licensed CC BY-NC-SA (non-commercial) and was therefore **not used**; we rebuilt from the raw sources, whose own terms govern.

Curation used SubPOP's published pipeline for question refinement and distribution computation, plus our own serializer for the final SFT set — byte-parity tested against SubPOP's prompt construction, with three integrity gates: every distribution must sum to 1.0 (±1%), numpy ≥ 2.0 scalar reprs are normalized by a whitelist regex before parsing (values identical; only the rendering changed between numpy versions), and any malformed cell is a loud stop, never a silent skip.

The result: **3,229 questions × 22 demographic subpopulations → 142,052 chat-format training examples**, with a question-level train/val split (no question appears in both). Each example uses the QA-steering format: one answered demographic multiple-choice item steers the subgroup, a bridge line, then the target survey item with an answer-forcing instruction; the completion is a **single option letter**. Hard labels are sampled from the empirical distribution (two samples per pair), so standard cross-entropy training reproduces the distribution in expectation — replicating SubPOP's `ce` objective. Training examples contain only subgroup-level aggregates re-sampled to letters; no individual-level records leave the curation environment.

3.2 Model and training

- Base: a **~9B open-weight instruct model** with a permissive commercial license, trained via LoRA on a hosted service.

- Configuration: **LoRA rank 8** (SubPOP's rank), **2 epochs**, platform-default learning rate.
- Outcome: final eval loss **0.5582** → perplexity **1.75** on held-out questions in the training format.

One observation was logged at deployment and deliberately left untreated: smoke-tested on a steered prompt *from its own training distribution* with the instruction to answer with a single letter, the served adapter answered **"CCC"** — the right letter, without termination. Per the pre-registration (§3.4), we added no stop sequences or serving-side accommodation before grading.

3.3 Evaluation protocol

The candidate was evaluated as the sole model behind every simulated respondent of its runs. Three properties keep the comparison honest:

1. **No fallback.** A pilot respondent whose call fails is marked degraded and *counted*, never silently answered by another model — otherwise a failing candidate is graded as a blend.
2. **Identical harness in both arms.** Same 40-item bank of real polling questions, same n = 100 simulated respondents per item, same persona generation, prompts, and parser, same session. Both arms were evaluated **raw**: the deployed product applies additional proprietary post-processing, which was disabled in both arms so the comparison isolates the model stacks under an identical elicitation harness.
3. **Quarantined readings.** Pilot results are excluded from public reporting by protocol; this article is their disclosure.

3.4 Pre-registration

Decided before any training:

- **Win:** pilot raw MAE < baseline raw MAE − 1.3pp. The 1.3pp band is the run-to-run noise measured in a prior A/B at n = 100.
- **Within the band:** inconclusive → stop; no iteration without a new hypothesis. **Worse:** stop and publish.
- **>10% degraded respondents:** stop and report as a format-transfer finding. No prompt-side workarounds — a workaround would grade the workaround, not the candidate.

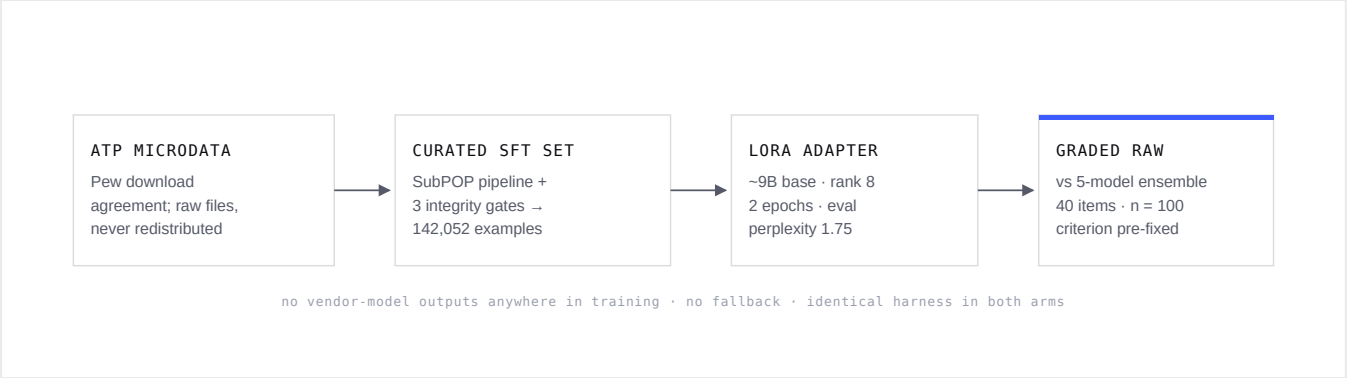

Figure 1 — From raw microdata to a graded pilot. The criterion (win = beat the ensemble by more than the 1.3pp noise band) was fixed before training began.

4 Results

Same 40 items, same day, n = 100 per item, both arms raw:

ARM (BOTH RAW)	MAE (PP) ↓	DIRECTIONAL PARITY ↑	PEARSON R ↑
Frontier ensemble	**12.53**	74.9%	0.709
Fine-tuned pilot	**46.92**	8.6%	−0.595

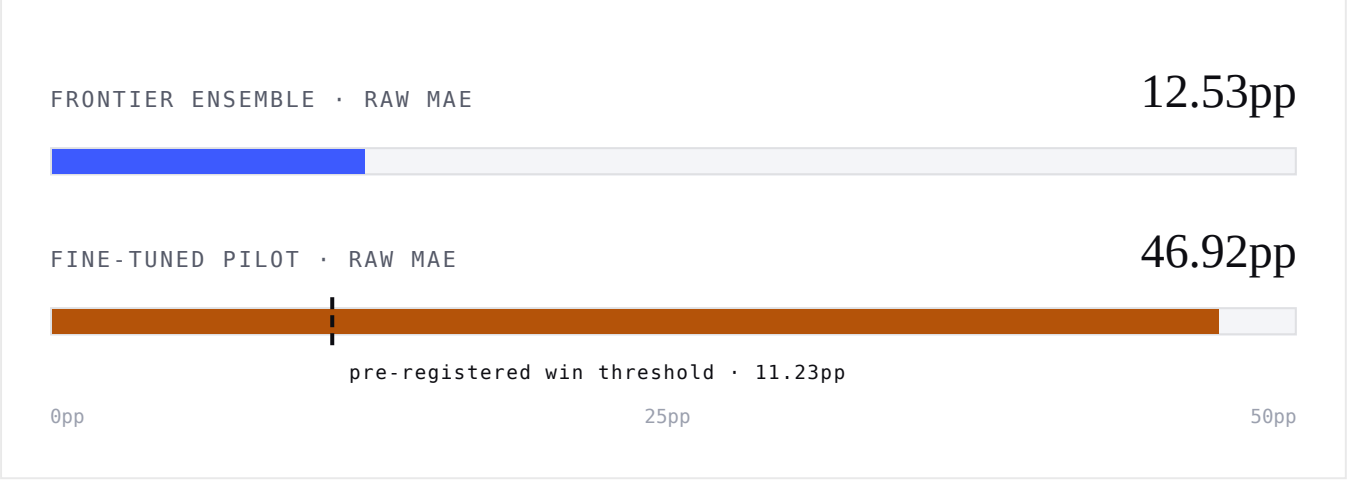

Figure 2 — The pre-registered criterion, as measured. The pilot needed to land left of the dashed threshold (baseline − 1.3pp noise band); it landed 34.4 points to the right of the baseline — 26× the noise band.

The pre-registered criterion required the pilot to land below **11.23**. It landed **34.4 points above the baseline** — and **16.3% of its seated respondents (636 of 3,900)** returned output the parser could not grade at all, versus nominal transient levels in the baseline arm.

Verdict: the pre-registered criterion was not met, by a margin 26× the noise band.

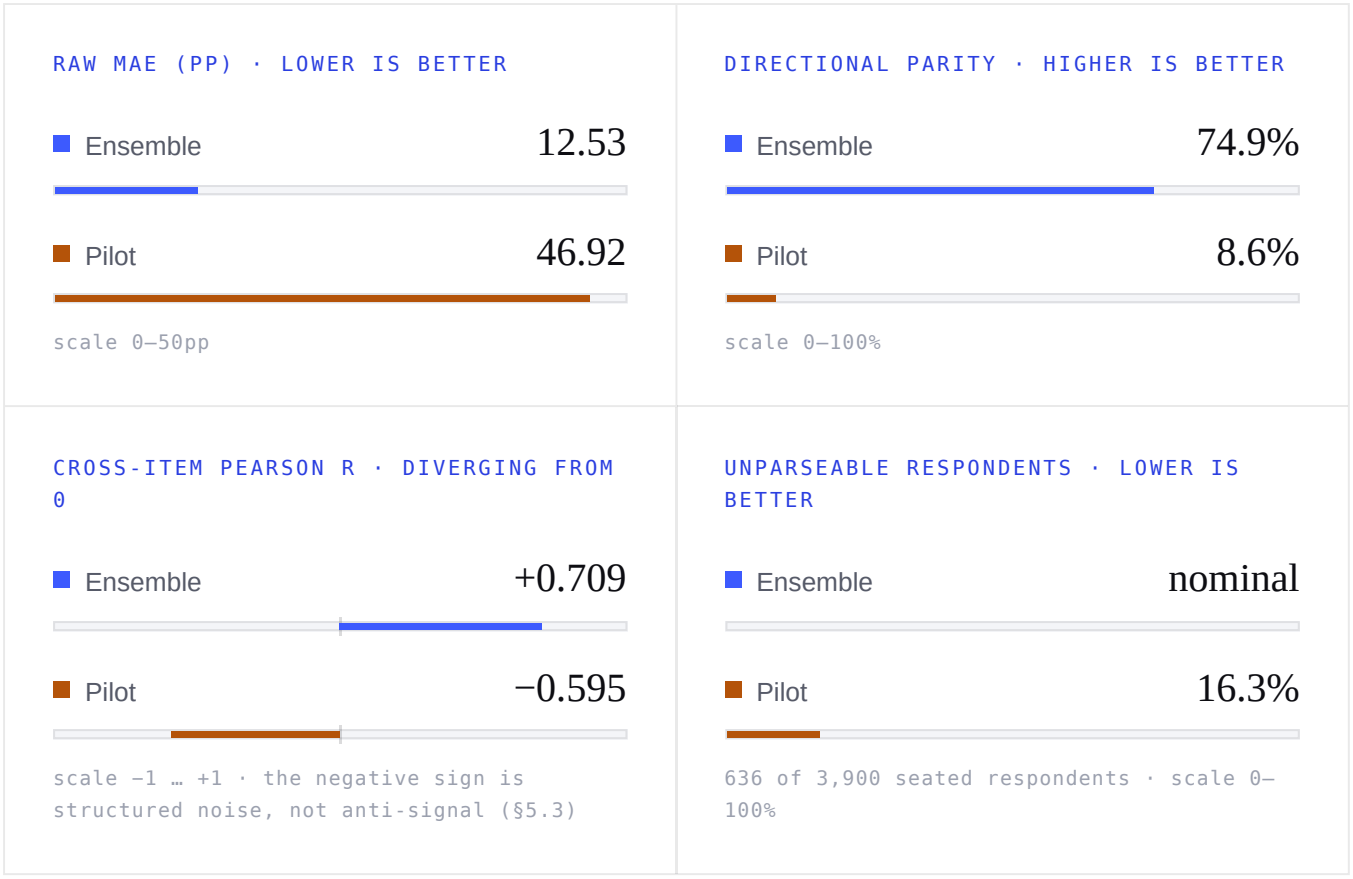

Figure 3 — The full result, on one board. Every metric is measured under the identical raw harness of §3.3; "nominal" reports the study's characterisation of transient baseline failures, not a computed rate.

5 Analysis: Anatomy of a Format-Transfer Failure

5.1 The adapter learned what it was taught

The most instructive number in this study is the one that looks best. An eval perplexity of 1.75 in the training format is close to the value implied by *faithfully reproducing* typical item distributions — for illustration, perfectly matching a 70/30 split yields perplexity $e^{0.611} \approx 1.84$ on the answer token — and far from degenerate mode-collapse behavior (asserting the majority option with near-certainty scores on the order of 4 on such items). The adapter did not memorize modes; it learned calibrated distributions over letters, which is precisely SubPOP's training objective.

Figure 4 — The number that looks best. Measured eval perplexity (1.75) sits where faithful distribution reproduction lives, far from mode-collapse territory: the adapter genuinely learned its objective. The landmark values are the illustrations of §5.1.

The metric was honest about the objective. The objective was not the job.

5.2 What the deployed task actually asks

A simulated respondent is not asked "which letter would your demographic pick." It receives a full persona and an open question, and must return a **structured first-person response** — an in-character answer with a stance and free-text reasoning, in a strict machine-parseable format that downstream aggregation depends on.

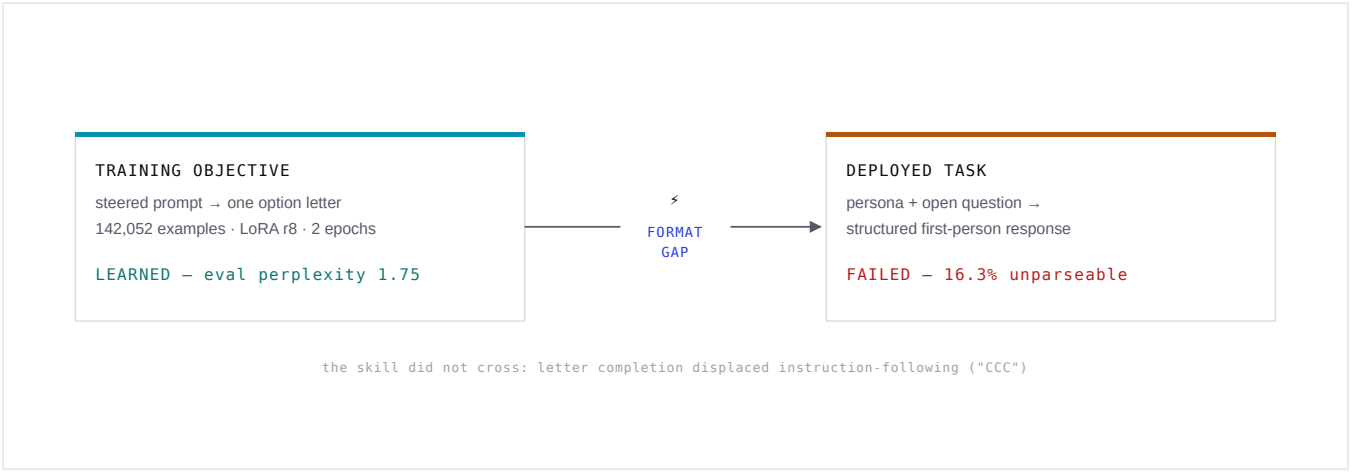

Figure 5 — The format-transfer failure. The adapter mastered its training objective and could not perform the deployed task; the narrow completion objective displaced the base model's instruction-following within the domain.

Nothing in 142,052 letter-completion examples teaches any of that. The adapter, asked for the structured response, emitted letter-like repetitions; asked for even a single letter in its own format, it emitted "CCC". Two epochs of single-token completions appear to have overwritten the base model's instruction-following and termination behavior within this domain — a

catastrophic-interference pattern familiar from the fine-tuning literature, here expressed as an inability to stop.

5.3 Reading the degenerate numbers

Two headline numbers deserve careful interpretation rather than face value:

- **16.3% degraded** is not an infrastructure failure; it is the candidate failing the format contract. The parser excludes and counts such respondents — it never invents a vote for them (a standing integrity rule: a response that cannot be parsed is excluded and reported, never defaulted).
- **r = −0.595** does not mean the model holds inverted opinions. When responses collapse toward position/letter patterns independent of item content, cross-item ranking becomes structured noise; interaction between a letter bias and the bank's option ordering is the parsimonious account of the negative sign. We did not fully root-cause it and treat it as *no signal*, not *anti-signal*.

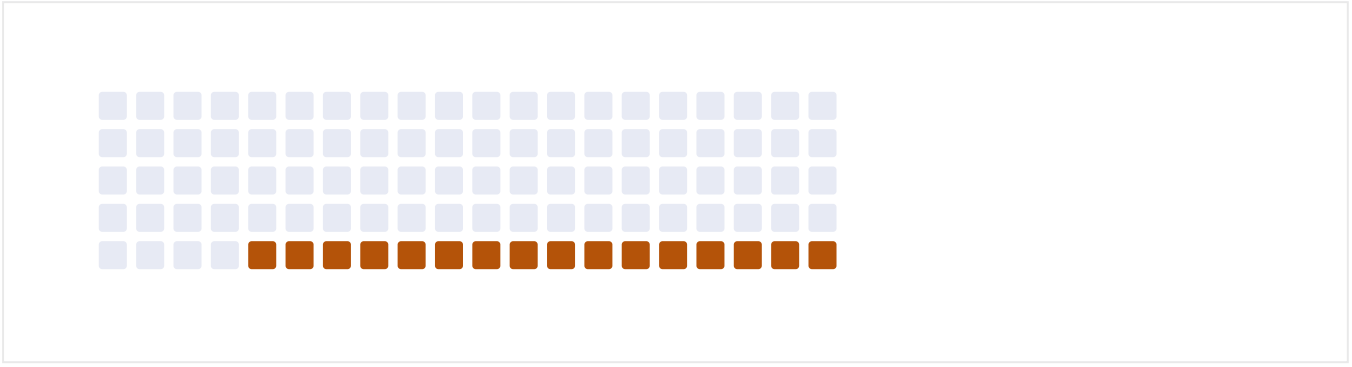

Figure 6 — The format contract, failed at scale: 636 of 3,900 seated pilot respondents (16.3%) returned output the parser could not grade. Each square ≈ 1% of the seated panel; the baseline arm ran at nominal transient levels. Unparseable respondents are excluded and counted — never defaulted to a vote.

5.4 Why we did not patch it

Stop sequences, format few-shots, constrained decoding, or a re-prompt wrapper would very likely have raised the parse rate. We declined, for a pre-registered reason: the criterion exists to grade the candidate, and every accommodation shifts the graded object toward the accommodation. There is also a deeper reason: even a perfectly terminated letter is not the task. The deployed task needs open-ended reasoning, within-persona consistency, and

answer-once discipline — capabilities the training signal did not merely fail to teach but actively displaced.

5.5 The finding

Distribution learning and respondent simulation are different tasks. Aggregate fidelity over subgroups (the Argyle/SubPOP line) and individual-level simulacra (the Park line) are not points on one axis, and gradient descent on the first does not purchase the second. For any system whose unit of value is the individual simulated respondent, a subgroup-marginal predictor is the wrong shape of artifact — even when it is *good at its own job*, which ours measurably was.

6 Discussion

6.1 Why a heterogeneous prompted ensemble is a hard baseline

- **Format robustness comes free.** Frontier instruct models arrive with strong instruction-following and termination discipline from broad training; the fine-tune demonstrably traded exactly those capabilities for its narrow objective.
- **Decorrelated priors.** Models from different providers err differently; aggregating across them attenuates idiosyncratic bias. A single tuned model concentrates one base's biases and adds its fine-tuning distribution's on top.
- **The baseline improves on its own.** A prompted ensemble inherits every upstream model improvement at zero retraining cost; a fine-tune is frozen at training time and depreciates against a moving frontier.

6.2 Implications for practice

In this market, "we trained our own model" is routinely presented as a mark of quality. This result argues the prior should run the other way until evidence is shown. Our fine-tune verifiably mastered its training objective — and still underperformed a prompted heterogeneous baseline by 34.4pp raw, while losing the ability to complete the deployed task format at all for a sixth of its respondents. Fine-tuning is not free accuracy: it trades general

capability for a narrow objective, and when the narrow objective is not *exactly* the deployed task, the trade can be catastrophic.

The comparison that settles it is cheap and specific: a pre-registered, same-harness, raw-versus-raw evaluation against a prompted frontier baseline on real ground truth. Any proprietary-model accuracy claim that arrives without one should be discounted accordingly.

7 Limitations

- One base model (~9B), one recipe, one training run — no seed variance, no scale sweep. The criterion answers "does *this* recipe beat the ensemble," not "can any fine-tune."
- Eval arms at n = 100 per item; the pre-registered ±1.3pp band reflects that scale. (The observed gap exceeds it by 26×.)
- We evaluated full substitution — every respondent on the candidate; mixed-ensemble configurations were out of scope.
- Serving-side termination settings were deliberately left untuned (§5.4); a format-matched fine-tune with tuned serving might parse — parsing was never the bar, beating 11.23pp was.
- The corpus and the benchmark are both US-centric; conclusions are scoped accordingly.
- The negative cross-item correlation is explained but not fully root-caused.

8 Future Work

The negative result closes a recipe, not the question. Directions we consider worth testing, in rough order of promise:

1. **Format-native corpora.** Programmatic conversion of human survey microdata into the deployed response format (human data, machine reshaped, with no vendor model in the loop), so the training objective and the task coincide.
2. **Two-stage training.** Distribution knowledge first, then a format-adaptation stage — testing whether the two skills can be stacked rather than traded.
3. **Constrained decoding**, evaluated honestly as a serving-time aid — with the caveat that it changes the graded object.

4. **Stronger open bases**, as open-weight instruction quality approaches the frontier; the failure mode in §5.2 may shrink with base robustness.
5. **Root-causing the negative correlation**, and extending the corpus and bank beyond the US.

9 Data, Licensing, and Ethics

- **No frontier-vendor model outputs** appear anywhere in training — provider terms prohibit it, and for a system evaluated against those models it would be methodologically circular.
- Raw microdata was handled under its sources' own terms: never redistributed, never committed to a repository; the training file contains subgroup-level aggregates re-sampled to hard labels, not individual records.
- The non-commercial packaged dataset was not used; restricted (not-for-profit) survey programs were excluded entirely, regardless of how permissively mirrors label them.
- The pilot system was never exposed to customers.
- We cannot release the training set (source terms), but the method is fully reproducible from the cited public pipeline plus one's own data grant; the serializer's integrity gates are described in §3.1.

10 Conclusion

We tested a live hypothesis from the literature — that fine-tuning on survey microdata transfers to respondent simulation — under a pre-registered criterion, and it failed decisively: 46.92pp versus 12.53pp raw error, with 16.3% of the fine-tuned panel unable to answer at all. The failure is precisely characterized: the adapter learned its distributions (perplexity 1.75) and lost the task (format transfer). Distribution learning is not respondent simulation.

The practical reading for this category: a task-specific fine-tune is a liability until it has beaten a prompted frontier baseline under the deployed task format, on real ground truth, with the criterion fixed in advance — and accuracy claims built on proprietary models should be held to that comparison. Negative results are underreported in this field; we publish this one in full because the category's claims currently outpace its published evaluations.

11 Addendum (August 28, 2026): Post-Publication Validity Audit

After publication, we identified two confounds the original evaluation had not isolated: the pilot was served at aggressive FP4 quantization while the baseline ran provider APIs at full precision, and no un-tuned base-model control was run. We audited both under a pre-registered protocol whose outcomes could refine the diagnosis but not reopen the verdict.

- **Quantization is exonerated.** Re-served at BF16 full precision, the adapter reproduces the failure exactly: it streams the answer letter as an unterminated repetition into the model's reasoning channel, leaving the answer channel empty — at temperature 0 and default alike. The pathology is the adapter's at any precision.
- **The reasoning-template mechanism is base-native.** The base model family serves with a thinking mode enabled by its default chat template, and the serving API offers no request-level off-switch. Serving stacks render this differently: the original evaluation's stack inlined the reasoning text into the answer channel — which is why 83.7% of pilot responses still parsed and were graded despite contamination — while dedicated serving routes it to a separate reasoning field, leaving the answer channel empty entirely.
- **The base-model control proved unmeasurable under available serving.** The bare, un-tuned base consumed a full 2,400-token budget in reasoning without ever emitting an answer on the serving configuration available to us, so a controlled base-vs-adapter comparison could not be produced. §7's open question about base viability remains open — now with a characterized cause.

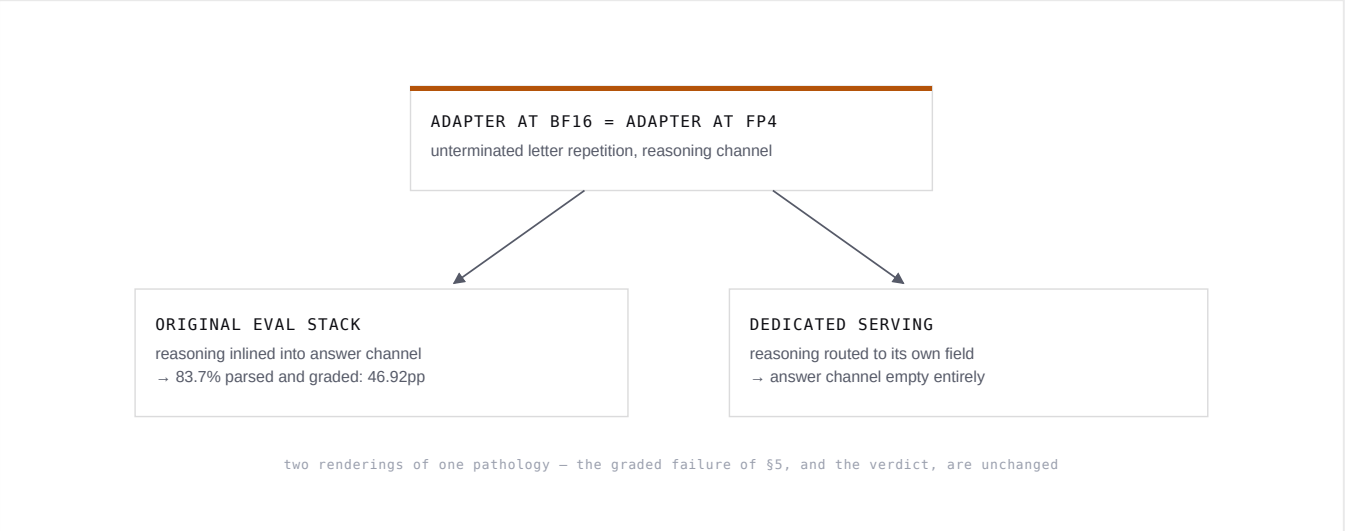

Figure 7 — Where the answer went. At any precision the adapter emits its learned letter as an unbounded repetition into the reasoning channel; serving stacks merely differ in where that

text surfaces. The audit refines the diagnosis without reopening the verdict.

The refinement leaves the finding two-layered rather than overturned: a base-native template behavior contaminated the output channel (and plausibly part of the 16.3% degraded fraction), while the graded failure — computed from the 83.7% of responses that parsed — remains the training-induced displacement described in §5. At full precision, the base produces structured reasoning; the adapter produces an unbounded repetition of its learned answer letter. One practice recommendation emerges for anyone running similar evaluations: smoke-test a candidate's template and termination behavior on the exact target serving stack before spending on benchmarks — it costs a fraction of a cent and predicts the parse-failure mode in advance.

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] 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*.

[4] Park, J. S., et al. (2024). Generative Agent Simulations of 1,000 People. *arXiv preprint*.

[5] Suh, J., et al. (2025). SubPOP: Fine-tuning Language Models on Scaled Survey Data for Predicting Distributions of Public Opinion. Code: github.com/josephjeesungsuh/subpop (BSD-3).

[6] CrowdOS benchmark ledger and methodology: crowdos.ai/benchmarks.

CITE THIS REPORT

```
@techreport{crowdos2026distribution,
  title       = {Distribution Learning Is Not Respondent Simulation:
                 A Pre-Registered Negative Result on Fine-Tuning for
                 Simulated Survey Respondents},
  author      = {{CrowdOS Research}},
  institution = {Motif Motion Pte. Ltd.},
  number      = {MM-TR-003},
  year        = {2026},
  month       = {8},
  address     = {Singapore},
  url         = {https://motifmotion.sg/papers/distribution-learning-respondent-si
  note        = {Working paper, v0.2; post-publication
                 validity audit appended 2026-08-28}
}
```